Data Mining –
Suche nach verborgenen Mustern

Dipl.-Inform. I. Boersch
FB Informatik und Medien
Mai 2013

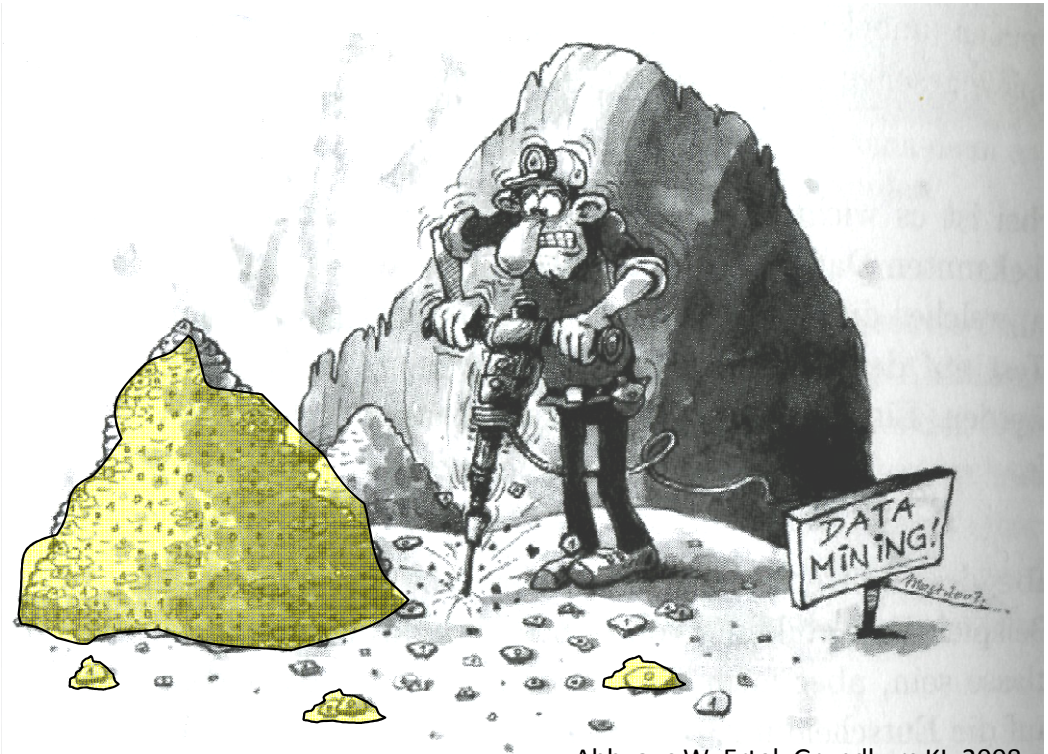


Abb. aus W. Ertel, Grundkurs KI, 2008

Firmen werben mit Data Mining

17. Mai 2014 - Meldung von N24: Eine Firma in Hong Kong hat ein Computerprogramm zum vollwertigen Vorstandsmitglied ernannt. "VITAL" soll durch künstliche Intelligenz unternehmerische Erfolge voraussagen und vor Pleiten warnen.

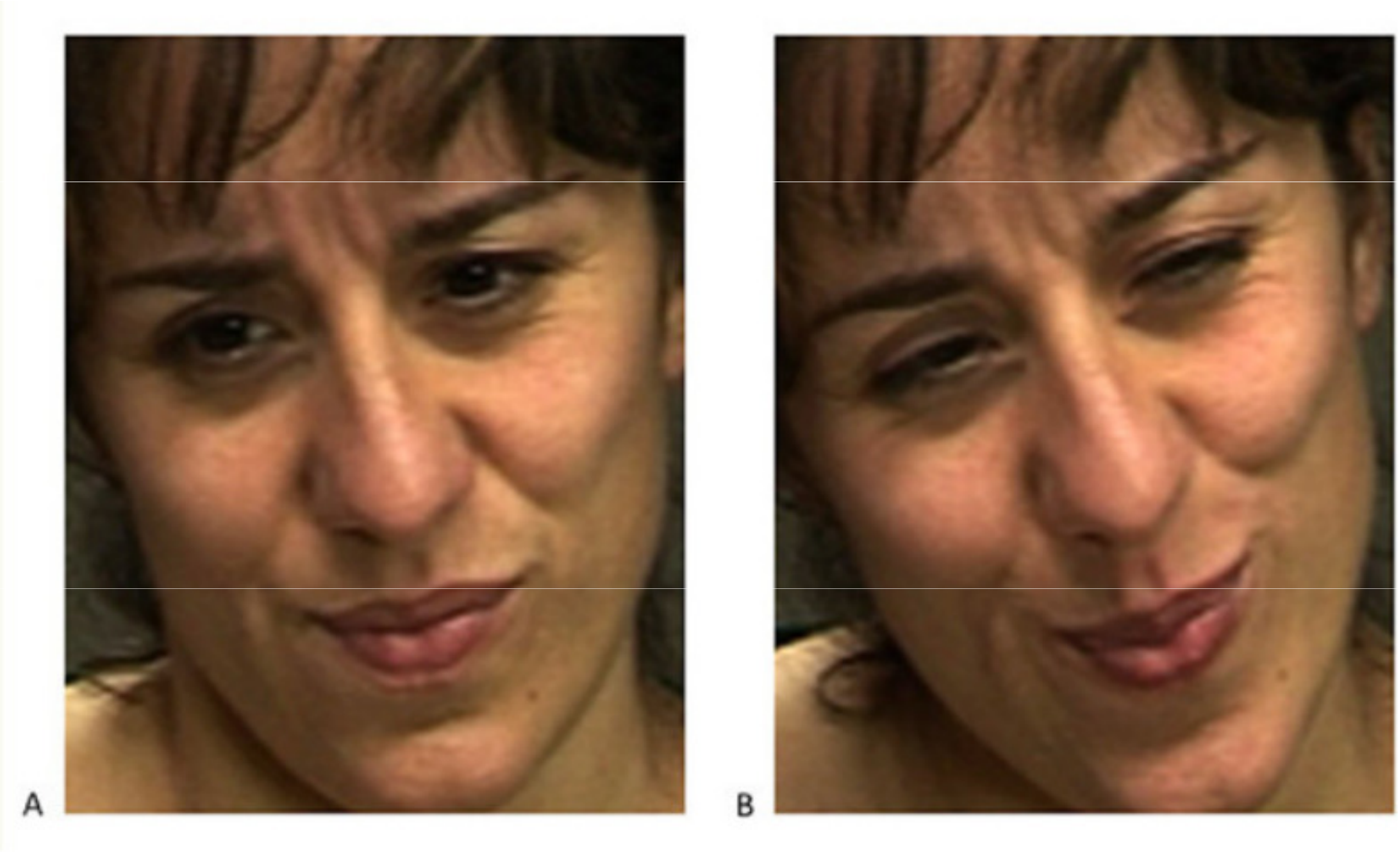
"Deep Knowledge Ventures" (DKV) ist eine chinesische Investmentfirma, die sich auf Projekte zu altersbedingten Medikamenten und regenerativer Medizin spezialisiert hat.. Das Besondere an DKV ist das neueste Vorstandsmitglied: "VITAL" ist ein von britischen Analysten programmierter Algorithmus – ein Programm mit extrem schneller Auffassungsgabe.
Indem es in kürzester Zeit große Mengen an Daten analysiert, kann es angeblich die besten Investitionsempfehlungen abgeben. Dazu scannt er die finanzielle Lage von Unternehmen, macht sich einen Überblick über klinische Versuche, geistiges Eigentum und den finanziellen Status des Kandidaten und gibt danach seinen Tipp ab.

"Unser Ziel ist es, Aufmerksamkeit zu erregen, indem wir ihn (VITAL) zu einem unabhängigen Entscheidungsträger machen“, sagte Charles Groome aus dem DKV-Vorstand dem "Business Insider". Bisher hat der Algorithmus, nach Groomes Angaben, schon zwei Beschlüssen zugestimmt. Wie clever dies war, wird sich aber erst langfristig zeigen.

VITAL's software was developed by UK-based Aging Analytics.

Quelle: http://www.n24.de/n24/Nachrichten/Netzwelt/d/4760222/computerprogramm-wird-vorstandsmitglied.html

Schmerz - gespielt oder echt?



Credit: Image courtesy of University of California - San Diego

31. März 2014

ELEKTRONIK PRAXIS

Künstliche Intelligenz

Computersystem unterscheidet echten von gespieltem Schmerz

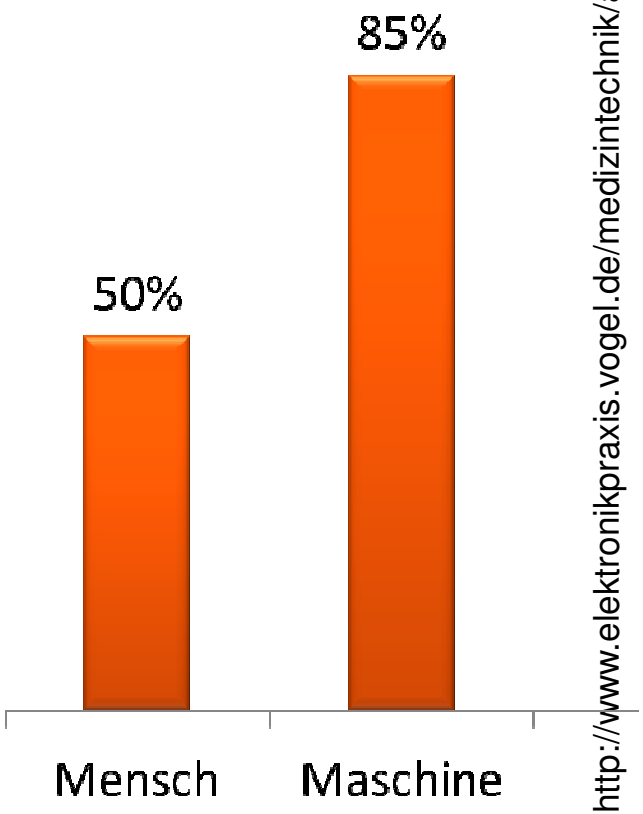
27.03.14 | Redakteur: Franz Graser

Der Kopfschmerz als Karikatur von George Cruikshank aus dem frühen 19. Jahrhundert. (Bild: Wikimedia Commons/Public Domain)

Forscher aus den USA und Kanada haben ein System entwickelt, das gespielte Emotionen im Gesichtsausdruck zuverlässiger erkennt als der durchschnittliche menschliche Beobachter. Bei Schmerz erreichte der Computer eine Erkennungsrate von 85 Prozent.

Das Studiendesign sah folgendermaßen aus: 25 freiwillige Versuchspersonen wurden jeweils zweimal gefilmt – einmal dabei, wie sie ihren Arm eine Minute lang in eisiges Wasser tauchen und dabei echten Schmerz empfinden und einmal, wie den Arm in lauwarmes Wasser

Erfolgsrate



Quelle: http://www.elektronikpraxis.vogel.de/medizintechnik/articles/439626/

Bartlett, M., Littlewort, G., Frank, M., & Lee, K. (2014). **Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions.** Current Biology, 2014 Mar 31; 24(7):738-43

Hausaufgabe: Rapidminer kennenlernen

- Wir verwenden in diesem Jahr die „Community Edition“ Rapidminer 5.3
- 5 Video Tutorials:

  0: Übersicht:
  http://www.youtube.com/watch?v=Cf5aNBvWYK0

  1: User Interface and First Process:
  http://www.youtube.com/watch?v=ABOSuFNYVTA

  2: Data Import and Repositories:
  http://www.youtube.com/watch?v=obmNNWglGGc

  3: Modeling and Scoring:
  http://www.youtube.com/watch?v=MfsqES7e1Es

  4: Visualization:
  http://www.youtube.com/watch?v=_8HzwQCFFfw

- Freiwillig: [Rapidminer download, ausprobieren]
  http://sourceforge.net/projects/rapidminer/
- Übung beenden

Übung

- Aufgabentypen des Data Mining
- K-means
- Support und Konfidenz von Regeln
- Zerlegung des Merkmalsraumes durch Entscheidungsbäume
- Entropie
- Erwartete Entropie

Ausgangsfrage

Ist es möglich, die angefallenen und weiter anfallenden riesigen Datenbestände in nützliche Informationen oder sogar Wissen umzuwandeln?

Falscher Alarm auf der Intensivstation

- TU Dortmund, Universitätsklinikum Regensburg, 2007
- Monitore überwachen den Zustand von Patienten
- Schwellwerte -> Alarm in kritischen Situationen

Problem: viele Fehlalarme

- Aufmerksamkeit sinkt, Schwellwerte werden erhöht, Messungen werden sogar deaktiviert

Aufgabe: Erkennen, wann ist ein Alarm wirklich wichtig.

- höchstens 2% echte Alarme verpassen (minimale Sensitivität = 0.98)

Ergebnis des Data Mining: 33% weniger Fehlalarme

[SG07] Sieben, W., Gather, U.: **Classifying Alarms in Intensive Care** - Analogy to Hypothesis Testing, in Springer's Lecture Notes of Computer Science Series: Artificial Intelligence in Medicine. Proceedings of the 11th **Conference on Artificial Intelligence in Medicine**, Vol. 4594/2007, eds. R. Bellazzi, A. Abu-Hanna, J. Hunter, Berlin / Heidelberg: Springer, 130-138, 2007

Übersicht

- Data Mining (DM) und Knowledge Discovery in Databases (KDD)

- Aufgabenstellungen des DM, Beispiele
- Wissensrepräsentation
- Entscheidungsbäume:
  - Repräsentation
  - Lernen / Konstruieren
  - Praktisch
- Performance von Klassifikatoren
- Ethik

Einordnung und Begriff KDD und DM

Knowledge Discovery in databases (KDD) *is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*

Data mining (DM) *is a step in the KDD process …*

[FPSS96]

gültige

neue

nützliche

verständliche Muster

- Im Sprachgebrauch häufig Gleichsetzung:
  - KDD = Data Mining im weiteren Sinne
  - Data Mining im engeren Sinne: Analyseverfahren

- Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic
- Definitionsüberblick in Wiedmann2001

Phasen des Knowledge Discovery in Databases (KDD)

Datenselektion / -extraktion

- Welche Daten notwendig und verfügbar?

Datenreinigung und Vorverarbeitung

- Fehlende Werte, Ausreißer, Inkonsistenzen,

Datentransformation

- Format für DM (einzelne Tabelle), Aggregation, Aufteilung in Trainings- und Testdaten

Data Mining im engeren Sinne (10 .. 90% Zeitaufwand je nach Datenqualität)

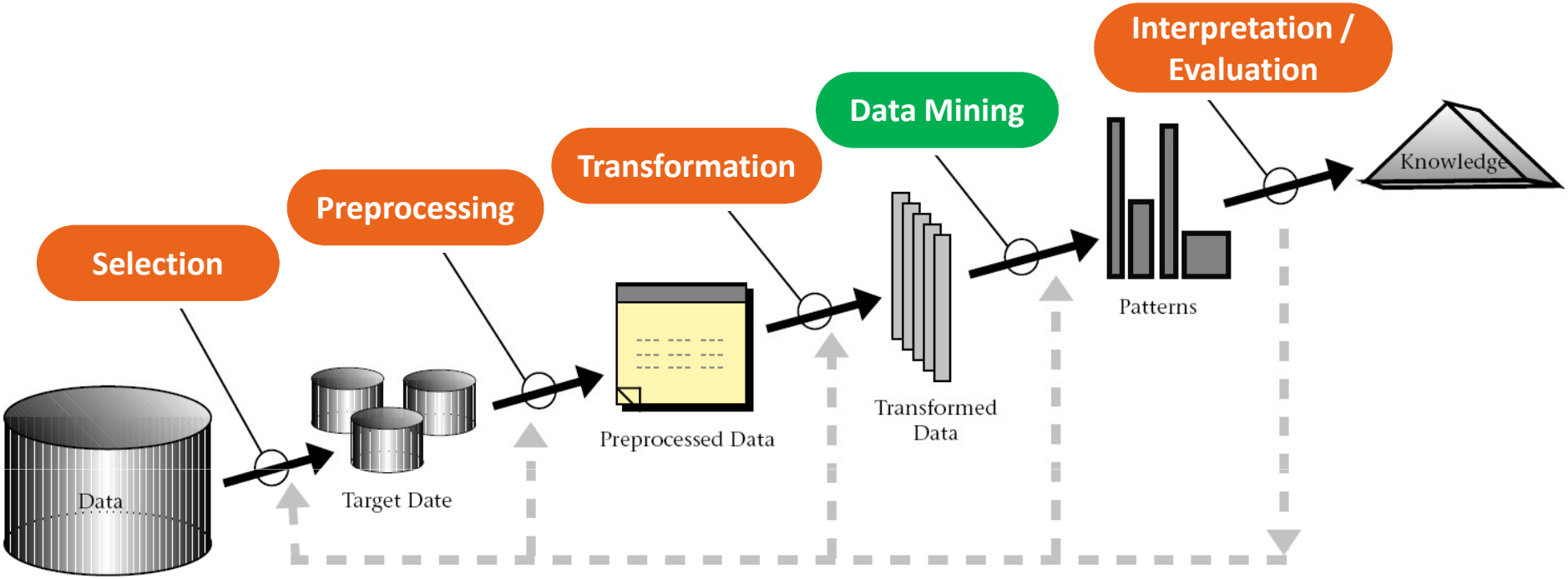
- Finden von Mustern: Exploration, Merkmalsextraktion und -selektion, Modellbildung

Interpretation und Evaluation

- Bewertung
- Präsentation
- Einsatz

Phasen des KDD

- Datenfluss mit Iterationen (Pfeile auch rückwärts)

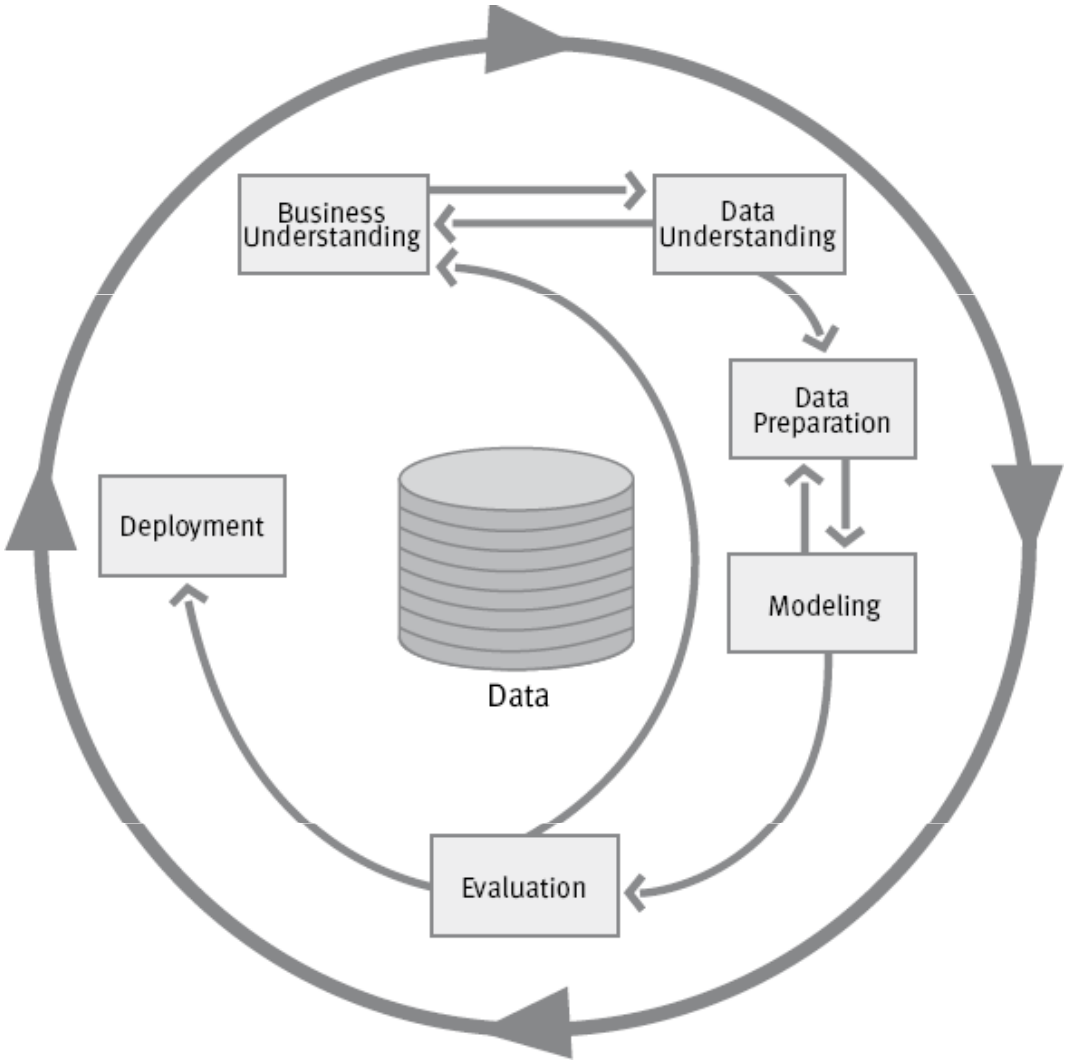


[FPSS96]

CRISP-DM reference model

- verbreitetes Modell zur praktischen Vorgehensweise, insbesondere **bei größeren Projekten**: Phasen und ihre Ergebnisse, Berichte, Akteure, Aufgabentypen ...

- Lebenszyklus eines Data Mining-Projektes besteht aus 6 Phasen
- keine feste Reihenfolge
- Hauptrichtung im Außenring als Zyklus



Empfehlung:
Chapman, Pete ; Clinton, Julian ; Kerber, Randy ; Khabaza, Thomas ; Reinartz, Thomas ; Shearer, Colin ; Wirth, Rudiger ; The CRISP-DM consortium (Hrsg.): **CRISP-DM 1.0 Step-by-step data mining guide**, 2000

Data Mining: Finden von Mustern

Ein Prozess (Unternehmen, Mensch, Maschine,...) erzeugt Daten

- Data Mining findet, konstruiert und optimiert **Modelle und Wissensrepräsentationen** zur **Beschreibung** und **Vorhersage** dieses Prozesses

Interdisziplinär: Statistik, Maschinelles Lernen, Mustererkennung, Datenbanken, ...

Querschnittstechnologie, d.h. unabhängig von einer Anwendungsdomäne

Ähnlich, aber zielgerichteter:

- **EDA (Exploratory Data Analysis):** Visualisieren, interaktiv Erforschen, keine klare Zielstellung -> Ergebnis: Formulieren von Hypothesen
- **OLAP (Online Analytical Processing):** Bestätigen von Hypothesen

Data Mining - Make sense from data

– Ähnliche Begriffe:
  - Biostatistics, Data Science, Machine learning, Signal processing, Business analytics, Econometrics, Statistical process control, Time Series, Analysis, Business Intelligence, Big Data, Predictive Analytics, Data-Driven Decision Making, Knowledge Extraction, Information Discovery, Information Harvesting, Data Archaeology, Data Pattern Processing.

Gemeinsame Fragestellungen, Prinzipien, Vorgehensweisen und Techniken.

Heute: Informatik immer beteiligt

Anwendungsbeispiele

- **Wie** erkennt der Händler, ob es sich bei einer Bestellung um einen zahlungswilligen Kunden handelt, der letztendlich die Ware auch bezahlt?
- **Wovon** hängt es bei der Evaluierung im FBI ab, ob ein Student in der Veranstaltung subjektiv „viel gelernt“ hat?
- **Wie** beeinflussen Einstellungen am Produktionsprozess die Qualität des Produktes?
- **Wie** erkennt man an der Dicke der menschlichen Netzhaut die Krankheit Multiple Sklerose?
- **Wie** viel Strom wird die Stadt Brandenburg morgen um 10 Uhr verbrauchen?…

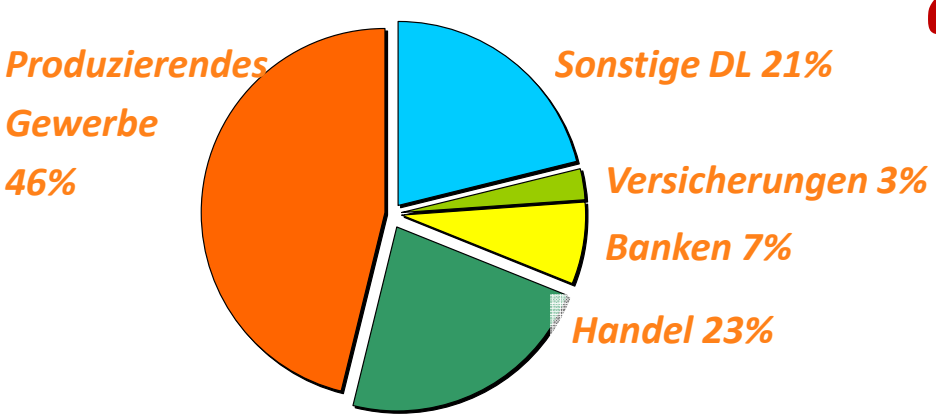
Beschreibung und Vorhersage

Beschreibung Vorhersage Beschreibung Vorhersage

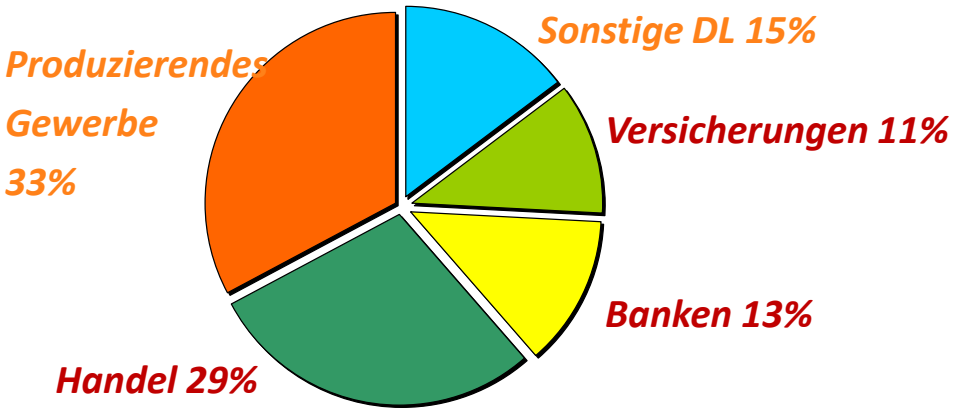
Analysieren Sie Ihre Kunden?

2002

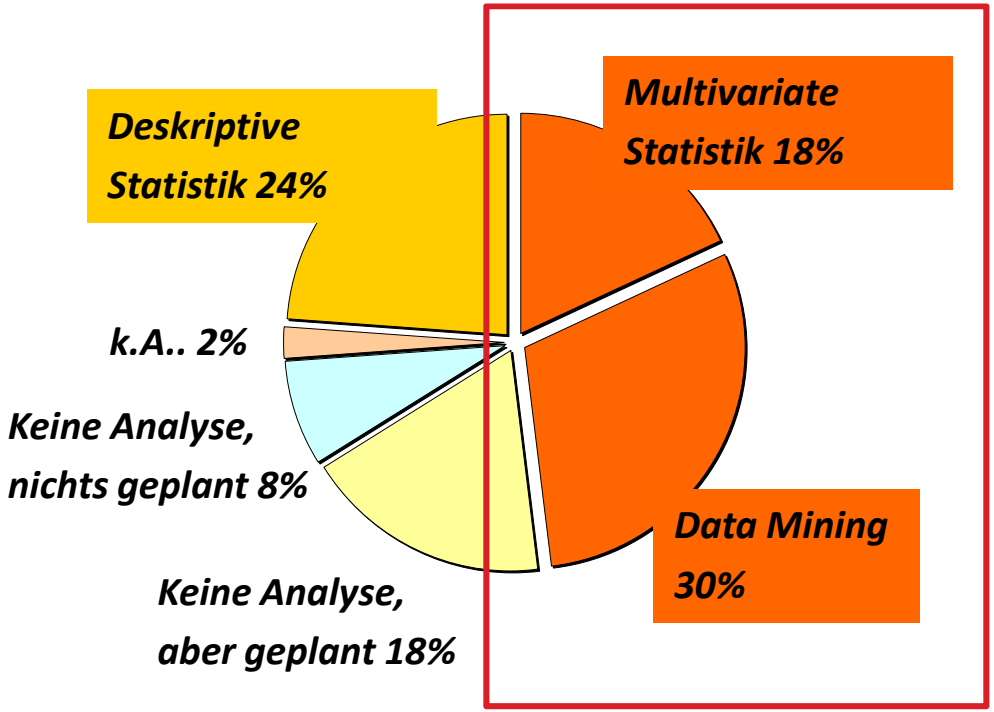
734 Unternehmen kontaktiert*



103 haben geantwortet



Art der Kundenanalysen:



* die 734 größten deutschen Unternehmen
Hippner, H., Merzenich, M. und Stolz, C. Data Mining: Anwendungspraxis in deutschen Unternehmen. In: Wilde, K.D., Data Mining Studie, absatzwirtschaft, 2002

2013

- **Gegenwart**:
  - dm: Personalplanung
  - otto: Verkaufsprognose
  - Vestas Wind Systems: wie viel Wind wird an einem Ort wehen?
- **Zukunft**: ständige Vorschläge
  - Verizon: Fernseher schlägt Paartherapie bei Streit vor, Patent
  - IAIS *: Menschenballungen, Katastrophen an Twittermeldungen erkennen.
  - Gesundheitskonzern Heritage. Wer muss im nächsten Jahr ins Krankenhaus?
  - EMI Music: Was wird ein Hit?

ZEIT ONLINE | INTERNET

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM

Internet | Datenschutz | Mobil | Games

ZEIT ONLIN

Partnersuche

BIG DATA

Wer hebt das Datengold?

Motorenhersteller, Supermärkte, Gebrauchtwagenhändler: Wie es Unternehmen inzwischen gelingt, aus der Informationsflut im Netz großen Profit zu schlagen.

Immer öfter wird die Frage gestellt: „Was sagen die Daten?“

Aber wenn man in diesen Tagen IT-Unternehmern, Unternehmensberatungsfirmen und manchem elektrisierten Konzernchef zuhört, bekommt man den Eindruck: Die Zeiten ändern sich, es ist ein neuer Goldrausch ausgebrochen. Die Pioniere von heute graben keine Flusslandschaften mehr um wie vor mehr als hundert Jahren am Klondike, sondern sie baggern in digitalen Datenbergen. Ihre Mine nennen sie Big Data – den großen Datenhaufen.

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) , Sankt Augustin

Predictive Policing in Chicago:

Schwarze Liste mit 400 Leuten, deren Verhalten, Eigenschaften oder Beziehungen auf **künftige** Verbrechen deuten - 60 bereits zuhause besucht

The minority report: Chicago's new police computer predicts crimes, but is it racist?

When the Chicago Police Department sent one of its commanders to Robert McDaniel's home last summer, the 22-year-old high school dropout was surprised. Though he lived in a neighborhood well-known for bloodshed on its streets, he hadn't committed a crime or interacted with a police officer recently.
And **he didn't have a violent criminal record, nor any gun violations.** In August, he incredulously told the Chicago Tribune*, "I haven't done nothing that the next kid growing up hadn't done*." Yet, there stood the female …

Kommentar von mattstroud, 19.02.2014

http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist

Income of Analytics/Data Mining/Data Science professionals

Jahreseinkommen in Dollar:

Region	Employment	2013 Avg. Salary	2012 Avg. Salary	% Change	2013 Count
US/Canada	**all**	**128.8**	**113.9**	**13.1%**	**223**
	Comp/Self	131.3	116.8	12.4%	194
	Univ/Gov	112.1	85.9	30.5%	29
Australia/NZ	**all**	108.1	111.8	-3.3%	8
	Comp/Self	112.9	108.3	4.2%	7
	Univ/Gov	75.0	127.5	na	1
W. Europe	**all**	85.1	78.1	8.9%	75
	Comp/Self	90.4	83.8	7.9%	62
	Univ/Gov	59.6	55.6	7.2%	13

Umfrage von http://www.kdnuggets.com am 26.02.2013 mit 383 Teilnehmern

Kaggle – Outsourcing von Data Mining

At the end of a competition, the competition host pays prize money in exchange for the intellectual property behind the winning model.

Competition Name	Reward	Teams	Deadline
Heritage Health Prize Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)	$3,000,000	1572	31 days
Predicting Parkinson's Disease Progression with Smartphone Data Can we objectively measure the symptoms of Parkinson's disease with a smartphone? We have the data to find out!	$10,000	0	24 days
Blue Book for Bulldozers Predict the auction sale price for a piece of heavy equipment to create a "blue book" for bulldozers.	$10,000	269	38 days
The Marinexplore and Cornell University Whale Detection Challenge Create an algorithm to detect North Atlantic right whale calls from audio recordings, prevent collisions with shipping traffic	$10,000	99	36 days
Job Salary Prediction Predict the salary of any UK job ad based on its contents	$6,000	119	25 days
Digit Recognizer Classify handwritten digits in this "Getting Started"			

Aufgabenstellungen des Data Mining

- DM und KDD, Phasen

- **Aufgabenstellungen des DM, Beispiele**
  - Klassifikation (häufigste)
  - Numerische Vorhersage
  - Abhängigkeitsanalyse
  - Clustering
  - Abweichungsanalyse

  - [Text-Mining]

1. Aufgabe: Klassifikation

Lernverfahren nimmt eine Menge klassifizierter Beispiele entgegen, aus denen es lernen soll, **unbekannte Beispiele** zu klassifizieren.

- Gegeben: Klassifizierte Stichprobe = die Trainingsmenge
- Gesucht: Modell zum Beschreiben und Vorhersagen von Klassen

- Transparente (= menschenlesbare) Modelle:
  - Entscheidungsbäume, Regelmengen, Bayessche Netze, Fuzzy-Systeme …
- Andere:
  - Neuronale Netze, logische Ausdrücke, Supportvektormaschinen RandomForrest, umfangreiche Modelle …

Tidy data – wie sollten Daten aussehen

1. Jedes Merkmal ist eine Spalte
2. Jede Beobachtung ist eine Zeile
3. Jede Tabelle beschreibt eine Art von Experiment.

	treatmentA	treatmentB
John Smith	—	5
Jane Doe	1	4
Mary Johnson	2	3

Table 1: Typical data presentation.

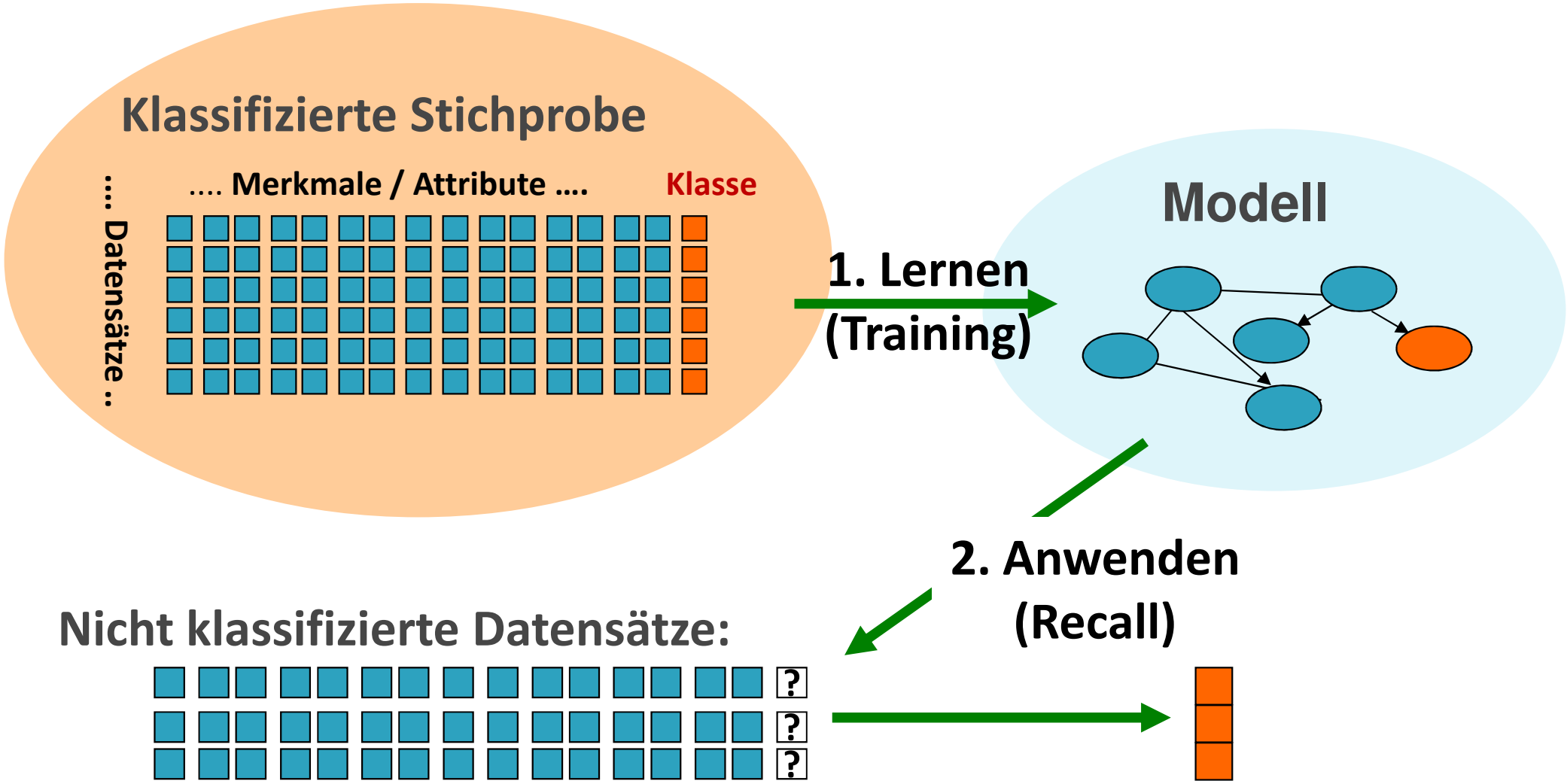
Merk-mal | Merk-mal | Klasse

name	treatment	result
Jane Doe	a	1
Jane Doe	b	4
John Smith	a	—
John Smith	b	5
Mary Johnson	a	2
Mary Johnson	b	3

Table 2: Tidied data.

[Tidy data. *The American Statistician* by Hadley Wickham. 2001 http://www.citeulike.org/user/Yanno/article/10869768]

2 Phasen der Klassifikation



Klassifikation – Beispiel (Data Mining Cup 2006)

Beschreiben

Vorhersagen

Vorhersage: Erzielt eine ebay-Auktion einen überdurchschnittlichen Verkaufserlös?

Stichprobe: 8.000 Online-Auktionen der Kategorie „Audio & Hi-Fi: MP3-Player: Apple iPod“

Merkmale: Titel, Untertitel, Zeitpunkte, Dauer, Rating, Startpreis, Sofortkaufen, Galerie, Fettschrift ...

Klasse: Hochpreis oder Niedrigpreis

- Anwenden auf 8000 unklassifizierte Auktionen -> Ergebnis einsenden, Ranking
- File dmc2006_train.txt

File dmc2006_train.txt

```
auct_id item_leaf_category_name listing_title   listing_subtitle        listing_start_date      listing_end_date        listing_durtn_days      li
00001   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple iPod mini 4 GB 2. Gen. 2005 silber blau pink grün         06/24/2005
00003   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       APPLE iPod mini Musicplayer 4GB Mac/Win USB 2.0 silber          08/13/2005
00004   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple iPOD Mini Silber NEUWARE mit Rechnung/Garantie            09/08/2005
00005   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple i-Pod mini 4GB BLUE  org. verpackt +Rechnung      Neuwertig +Garantie über (
00006   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple iPod mini 4 GB Rosa               09/18/2005      10/13/2005      25      7
00007   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple iPod mini 4 GB Rosa               09/19/2005      10/14/2005      25      7
00008   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple IPod mini blue mit Netzteil - absolut neuwertig           09/21/2005      1(
00009   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün iPod Mini 4GB (Grün) + Mega Zubehör und Dock + Rechnung ca. 1 Jahr alt deswegen no
00014   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün * Neu & OVP * i Pod mini 4GB Grün 1000 Songs PC+Mac             09/22/2005      1(
00015   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau iPod mini blau 4GB wie NEU mit Garantie und Socke OVP           09/22/2005      1(
00018   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       iPod mini silber, nagelneu!!!           09/22/2005      10/02/2005      1(
00020   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       ipod mini 4 GB silber wie neu Modell 2005 mit Rechnung  incl. Zubehör,netw
00022   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple iPod mini 4 GB blau - 1000 songs PC + MAC  OVP!!!         09/22/2005      1(
00024   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple ipod mini 2005  pink 4 GB wie NEU OVP             09/23/2005      10/03/2005
00025   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple iPOD Mini 4GB Blau - Neu mit Rechnung             09/23/2005      10/03/2005
00026   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple iPod mini 4 GB silber Mac/Windows dt. 2005 2.Gen.         09/23/2005
00027   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       ipod mini 4GB mit Zubehör !!!           09/24/2005      10/01/2005      7
00029   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple iPod mini 4GB mit Zubehör und itrip               09/23/2005      10/03/2005
00031   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple iPod Mini Pink 4GB / 2. Generation                09/24/2005      10/01/2005
00032   Audio & Hi-Fi:MP3-Player:Apple iPod:40 GB       **Apple i Pod Photo 40GB**TOP-Zustand**         09/24/2005      10/01/2005      7       1
00035   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün I-Pod Mini (Super Preis)        So gut wie neu  09/24/2005      10/01/2005      7
00036   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Ipod mini 4 GB, Silber - TOP ZUSTAND            09/25/2005      10/02/2005
00037   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       APPLE iPod mini 4 GB - 1000 Songs PC + Mac              09/24/2005      1(
00039   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau I Pod mini  4 GB  Blue Gerät  Neu               09/24/2005      10/01/2005      7
00040   Audio & Hi-Fi:MP3-Player:Apple iPod:40 GB       Apple iPod! 40 GB! 3. Gen.! Wie neu! Top gepflegt!              09/24/2005      10/04/2005
00043   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink *** Ipod Mini 4GB rosa***               09/24/2005      10/01/2005      7       1
00045   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün Apple iPod mini, 4GB - grün, super gepflegt            09/24/2005      10/01/2005
00046   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün iPod Mini 4GB           09/25/2005      10/02/2005      7       1       29      1.
00048   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple iPod mini pink - neu + Rechnung           09/25/2005      10/02/2005      7
00049   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Gold iPod mini m. Dock und Ledertasche!! ->für Salsa-Freunde         09/26/2005      1(
00052   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple iPod mini blau 4GB ***wie NEU*** in OVP           09/25/2005      10/02/2005
00061   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       iPod mini - Silber - neu                09/25/2005      10/02/2005      7
00063   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Gold NEU Apple iPod mini 4GB Gold NEU                09/25/2005      10/02/2005      7
00064   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       +++ iPod mini 4 GB silber neues Modell Garantie +++             09/25/2005
00065   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple Ipod mini 4 gb *Neu* & #OVP# 2 Jahre Garantie             09/25/2005
00066   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple iPod mini *silber* mit viel Zubehör               09/25/2005      1(
00067   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple IPod Mini 4 GB Silver Neu und Garantie            09/25/2005      1(
00070   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau ipod mini in blau - NEU und original verpackt!!!              09/25/2005      1(
00071   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple Ipod mini 4GB blau OVP            09/25/2005      10/02/2005      7       1
00072   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple iPod mini pink 4 GB mit Zubehör OVP Top Zustand          09/25/2005      1(
00073   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün I pod mini *neu*                09/25/2005      10/02/2005      7       1       1(
00074   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       Apple iPod Mini 4GB Silber NEU & OVP            09/25/2005      10/02/2005
00076   Audio & Hi-Fi:MP3-Player:Apple iPod:40 GB       Apple iPod Photo 40 GB - Top Zustand, OVP               09/25/2005      10/02/2005      7
00079   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau Apple i-Pod mini                09/25/2005      10/02/2005      7       1       2:
00080   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       iPod mini 4 GB silber originalverpackt          09/27/2005      10/02/2005
00081   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       iPod Mini Silber 4 GB mit Zubehör, Wie NEU!!            09/26/2005      1(
00084   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau iPod Mini 4 GB blau + ITrip FM Transmitter + OVP              09/26/2005      1(
00085   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink Apple IPOD mini pink neu 4 Gigabyte Mp3 Player itunes          09/26/2005      1(
00086   Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber       MINI-IPOD 4 GB Silver, OVP+Rechnung             09/26/2005      10/06/2005
```

File dmc2006_train.txt

Klasse

auct_id	item_leaf_category_name	listing_title	listing_subtitle	listing_start_date	listing_end_date	gms_greater_avg		
1	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Apple iPod mini 4 GB 2. Gen. 2005 silber blau pink grün		06/24/2005	11/03/2005	1		
3	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	APPLE iPod mini Musicplayer 4GB Mac/Win USB 2.0 silber		08/13/2005	10/11/2005	1		
4	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Apple iPOD Mini Silber NEUWARE mit Rechnung/Garantie		09/08/2005	10/23/2005	1		
5	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	Apple i-Pod mini 4GB BLUE org. verpackt +Rechnung	Neuwertig +Garar	10/14/2005	10/24/2005	1		
6	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple iPod mini 4 GB Rosa		09/18/2005	10/13/2005	1		
7	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple iPod mini 4 GB Rosa		09/19/2005	10/14/2005	1		
8	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	Apple IPod mini blue mit Netzteil - absolut neuwertig		09/21/2005	10/01/2005	1		
9	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün	iPod Mini 4GB (Grün) + Mega Zubehör und Dock + Rechnung	ca. 1 Jahr alt des	09/21/2005	10/01/2005	1		
14	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün	* Neu & OVP * i Pod mini 4GB Grün 1000 Songs PC+Mac		09/22/2005	10/02/2005	0		
15	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	iPod mini blau 4GB wie NEU mit Garantie und Socke OVP		09/22/2005	10/02/2005	0		
18	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	iPod mini silber	nagelneu!!!		09/22/2005	167977	0	
20	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	ipod mini 4 GB silber wie neu Modell 2005 mit Rechnung	incl. Zubehör	networx silicon skin	09/22/2005	167977	0	
22	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	Apple iPod mini 4 GB blau - 1000 songs PC + MAC OVP!!!		09/22/2005	10/02/2005	1		
24	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple ipod mini 2005 pink 4 GB wie NEU OVP		09/23/2005	10/03/20	0		
25	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	Apple iPOD Mini 4GB Blau - Neu mit Rechnung		09/23/2005	10/03/20			
26	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Apple iPod mini 4 GB silber Mac/Windows dt. 2005 2.Gen.		09/23/2005	10/03/20			
27	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	ipod mini 4GB mit Zubehör !!!		09/24/2005	10/01/20			
29	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple iPod mini 4GB mit Zubehör und itrip		09/23/2005	10/03/20			
31	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple iPod Mini Pink 4GB / 2. Generation		09/24/2005	10/01/2005	0		
32	Audio & Hi-Fi:MP3-Player:Apple iPod:40 GB	**Apple i Pod Photo 40GB**TOP-Zustand**		09/24/2005	10/01/2005	1		
35	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün	I-Pod Mini (Super Preis)	So gut wie neu	09/24/2005	10/01/2005	0		
36	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Ipod mini 4 GB	Silber - TOP ZUSTAND		09/25/2005	167977	0	
37	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	APPLE iPod mini 4 GB - 1000 Songs PC + Mac		09/24/2005	10/04/2005	0		
39	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	I Pod mini 4 GB Blue Ge			/01/2005	1		
40	Audio & Hi-Fi:MP3-Player:Apple iPod:40 GB	Apple iPod! 40 GB! 3. Ger			/04/2005	1		
43	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	*** Ipod Mini 4GB rosa***			/01/2005	0		
45	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün	Apple iPod mini				1 151.00	164757	0
46	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Grün	iPod Mini 4GB		09/25/2005	10/02/2005	0		
48	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple iPod mini pink - neu + Rechnung		09/25/2005	10/02/2005	1		
49	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Gold	iPod mini m. Dock und Ledertasche!! ->für Salsa-Freunde		09/26/2005	10/03/2005	0		
52	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	Apple iPod mini blau 4GB ***wie NEU*** in OVP		09/25/2005	10/02/2005	0		
61	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	iPod mini - Silber - neu		09/25/2005	10/02/2005	1		
63	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Gold	NEU Apple iPod mini 4GB Gold NEU		09/25/2005	10/02/2005	0		
64	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	+++ iPod mini 4 GB silber neues Modell Garantie +++		09/25/2005	10/02/2005	0		
65	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Apple Ipod mini 4 gb *Neu* & #OVP# 2 Jahre Garantie		09/25/2005	10/02/2005	1		
66	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Apple iPod mini *silber* mit viel Zubehör		09/25/2005	10/05/2005	1		
67	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	Apple IPod Mini 4 GB Silver Neu und Garantie		09/25/2005	10/02/2005	0		
70	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	ipod mini in blau - NEU und original verpackt!!!		09/25/2005	10/02/2005	0		
71	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Blau	Apple Ipod mini 4GB blau OVP		09/25/2005	10/02/2005	0		
72	Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	Apple iPod mini pink 4 GB mit Zubehör OVP Top Zustand		09/25/2005	10/02/2005	0		
						0		
						0		
7995	15992 Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Pink	APPLE iPod mini 4GB PINK MP3 Neu OVP		03/18/2006	03/19/2006	221153	1	
7996	15993 Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Mini 4 GB:Silber	---->Ipod mini SIlber 4GB TOP ZUSTAND<-----		03/18/2006	03/19/2006	1		
7997	15995 Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Nano 4 GB	Apple - iPod NANO 4GB(4 GB) schwarz /NEU/OVP/ Rechnung		03/18/2006	03/19/2006			
7998	15996 Audio & Hi-Fi:MP3-Player:Apple iPod:iPod Nano 4 GB	Apple - iPod NANO 4GB(4 GB) schwarz /NEU/OVP/ Rechnung		03/18/2006	03/21/2006			
7999	15997 Au	Apple - iPod NANO 4GB(4 GB) schwarz /NEU/OVP/ Rechnung		03/18/2006	03/23/2006			
8000	15998 Au	iPod 4.Generation 40GB Defekt fuer bastler		03/19/2006	03/22/2006			
8001	16000 Au	Apple - iPod NANO 2GB **schwarz** NEU mit Rechnung!		03/19/2006	03/22/2006			

Formatproblem

Textanalyse hilfreich?

DatenMENGE

DM-Cup 2008 - Kündigerprävention

121. Süddeutsche Klassenlotterie: Wer kündigt wann?

Beschreiben
Vorhersagen

Trainingsdaten: 113.477 Datensätze

- **Personendaten** (Alter, Geschlecht)
- abgeleitete Personendaten (Bankart, Telefonart)
- Marketingdaten (Werbeweg, Responseweg)
- Spieldaten (gewünschter Einsatz, div. Spielparameter)
- Kontaktinformationen (Kategorien: Information, Reklamation)
- 50 mikrogeografische Variablen, wie z.B. Altersverteilung, Kfz-Typen und -verteilung, Kaufkraft, Gebäudetypen, Konsumaffinitäten.

Klasse:

- Bezahlt gar nicht
- Bezahlt nur die 1. Klasse
- Bezahlt bis einschließlich 2. Klasse
- Bezahlt bis einschließlich 6. Klasse
- Bezahlt mind. bis einschließlich 1. Klasse der Folgelotterie

Rückblick DMC Wettbewerb 2008
Anmeldungen: 618
Beteiligte Universitäten: 164
Beteiligte Länder: 42
Eingereichte Lösungen: 231

DM-Cup 2013 – Kaufprognose

Ein Webshop beobachtet seine Besucher (überwiegend Frauen):
Kauft sie oder kauft sie nicht?

Trainingsdaten: 50.000 Sessions

- Kundendaten: Alter, Adresscode, Zahlungsanzahl, Acountlebensdauer, Datum der letzten Bestellung, Score, ...
- Session: Dauer der Session, Preisverlauf angesehener Artikel, Preise in den Warenkorb gelegter Artikel, Zustandsverlauf im Bestellprozess, ...

Klasse:

- Session endet mit Bestellung
- Session endet ohne Bestellung

Ziel: Automatisch während der Session Rabatte oder Up-Selling anbieten

99 Teams, 77 Hochschulen, 24 Länder, 79 Lösungen

FHB: Platz 22

DM-Cup 2014 – Retourenprognose

Ein Versandhändler, will beim Bestellen wissen:
Welcher Artikel der Bestellung wird zurückgeschickt?

Trainingsdaten: 481.092 bestellte Artikel

- Kundendaten: customerID, salutation, dateOfBirth, state, creationDate
- Bestellung: orderDate,
- Artikel: orderItemID, deliveryDate, itemID, size, color, manufacturerID, price

Klasse:

- Artikel zurückgeschickt
- Artikel nicht zurückgeschickt

Studenten des Data Mining-Projektes der FHB unter den besten 9!

125 Teams, 99 Hochschulen, 28 Ländern, 57 Lösungen
Mehr dazu bei Ihren Kommilitonen im Data Mining-Projekt
Platzierung bekannt am 2.-3.Juli 2014

Herzinfarkterkennung in der Notaufnahme

Beschreiben d Vorhersagen

Woran erkennt man einen Herzinfarkt (MI = myocardial infarction)?

Stichprobe: Patienten mit Brustschmerz in Notaufnahme in *Edinburgh (n=*1252) *und Sheffield (n=500)*

45 Merkmale: age, smoker, ex-smoker, family history of MI, diabetes, high blood pressure, lipids, retrosternal pain, chest pain major symptom, left chest pain, right chest pain, back pain, left arm pain, right arm pain, pain affected by breathing, postural pain, chest wall tenderness, sharp pain, tight pain, sweating, shortness of breath, nausea, vomiting, syncope, episodic pain, worsening of pain, duration of pain, previous angina, previous MI, pain worse than prev. Angina, crackles, added heart sounds, hypoperfusion, heart rhythm, left vent. hypertrophy, left bundle branch block, ST elevation, new Q waves, right bundle branch block, ST depression, T wave changes, ST or T waves abnormal, old ischemia, old MI, sex

Die Forscher erstellen einen Entscheidungsbaum

[TFLK98] Tsien, C., Fraser, H., Long, W. and Kennedy, R. Medinfo. v9. 493-497., 1998.: Using classification tree and logistic regression methods to diagnose myocardial infarction, online unter: http://groups.csail.mit.edu/medg/people/hamish/medinfo-chris.pdf

Entscheidungsbaum Herzinfarkterkennung

```
ST elevation = 1: 1
ST elevation = 0:
| New Q waves = 1: 1
| New Q waves = 0:
| | ST depression = 0: 0
| | ST depression = 1:
| | | Old ischemia = 1: 0
| | | Old ischemia = 0:
| | | | Family history of MI = 1: 1
| | | | Family history of MI = 0:
| | | | | age <= 61 : 1
| | | | | age > 61 :
| | | | | | Duration of pain (hours) <= 2 : 0
| | | | | | Duration of pain (hours) > 2 :
| | | | | | | T wave changes = 1: 1
| | | | | | | T wave changes = 0:
| | | | | | | | Right arm pain = 1: 0
| | | | | | | | Right arm pain = 0:
| | | | | | | | | Crackles = 0: 0
| | | | | | | | | Crackles = 1: 1
```

Nur 10 Merkmale, von bisher 45 werden für eine Schnelldiagnose benötigt!

Wie gut ist der Baum?

Sensitivity = 81.4%

Specificity = 92.1%

PPV = 72.9%

Accuracy = 89.9%

$$sens = \frac{TP}{TP+FN} \quad spec = \frac{TN}{TN+FP}$$

2. Aufgabe: Numerische Vorhersage

- Variante der Klassifikation
- Das vorhergesagte Ergebnis ist keine diskrete Klasse, sondern eine numerische Größe, also eine Zahl.

- Modelle:
  - Regression (bspw. lineare Regression)
  - Formeln und ihre Parameter, bspw. mit EA wie GA und GP
  - Regressionsbäume, Modellbäume
  - Zusätzlich: Modelle der Klassifikation, wie Entscheidungsbäume, neuronale Netze usw.

Formelausdrücke aus Daten: Erinnern Sie sich an die unbekannten Funktionen beim GP-Applet?

Lineare Regression –
Methode der kleinsten Quadrate (MKQ)

Annäherung der Trainingsdaten durch die Gerade mit kleinstem quadratischen Fehler

Koeffizienten dieser Geraden lassen sich berechnen!

$\hat{y} = mx + n = \mathrm{w}_1 \mathrm{x} + \mathrm{w}_0$ Regressionsgerade schätzt y

$E = \sum (y_i - \hat{y}_i)^2$ Quadratischer Fehler der Schätzung

$E = \sum (y_i - w_1 x_i - w_0)^2 -> \text{soll minimal werden}$

Nullsetzen der Ableitungen führt zur Gerade mit dem kleinsten quadratischen Fehler

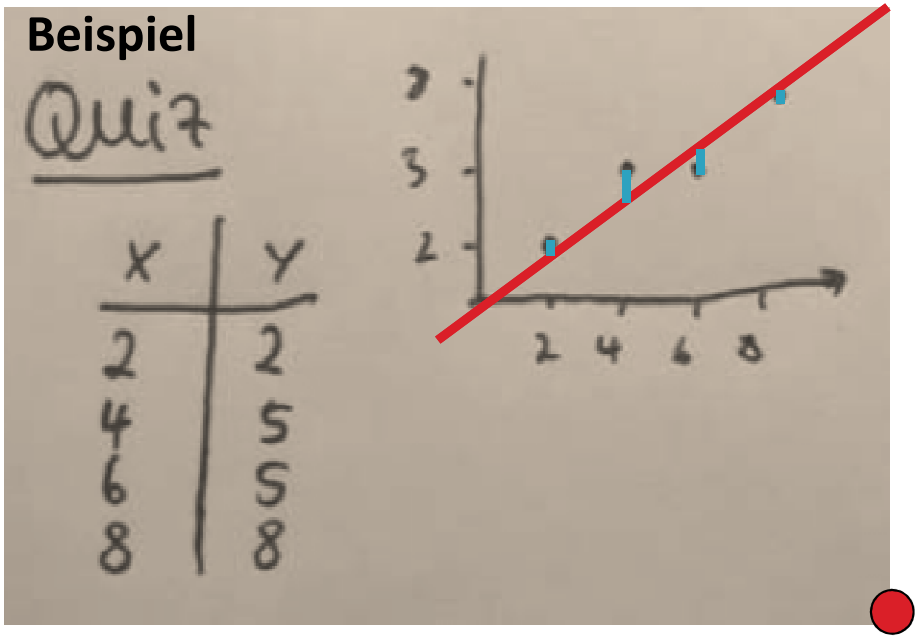
$\frac{\partial E}{w_0} \overset{!}{=} 0$ und $\frac{\partial E}{w_1} \overset{!}{=} 0$

...

$$w_0 = \frac{1}{N} \sum y_i - \frac{w_1}{N} \sum x_i$$

$$w_1 = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2}$$

Beispiel



Beispiel: Berechnung der Regressionsgeraden

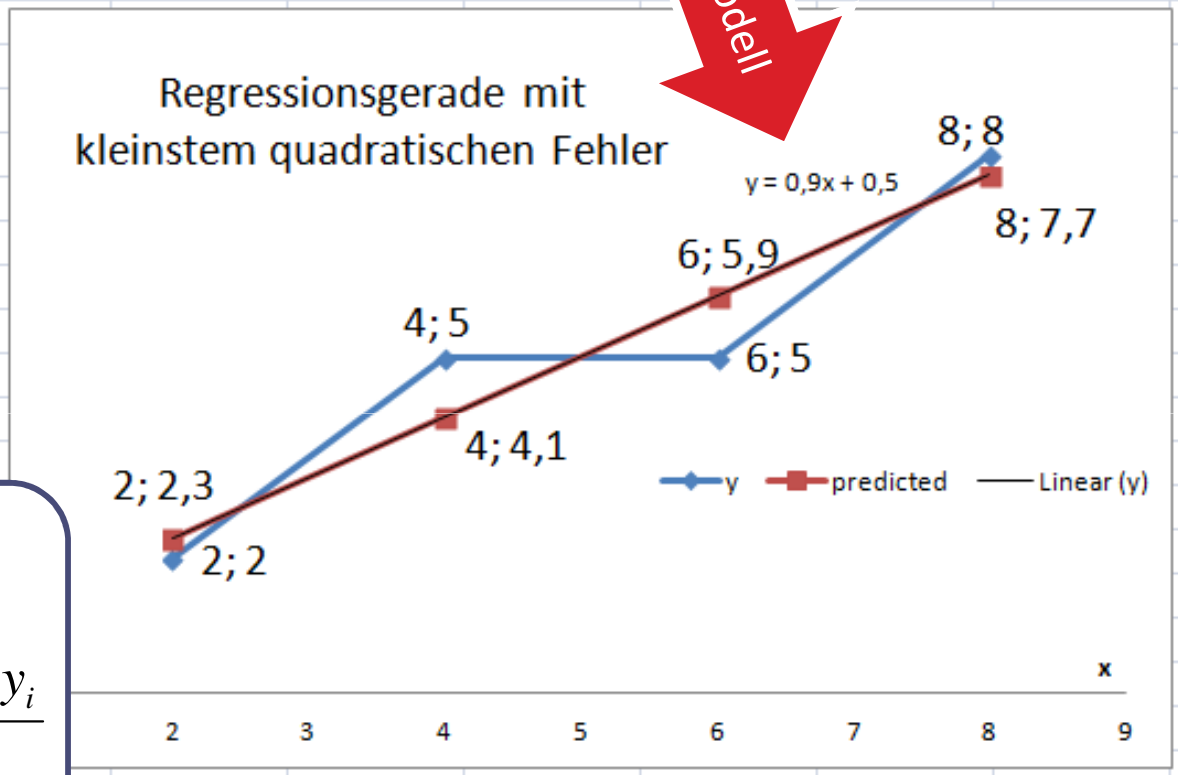
Aus acht Zahlen (Trainingsmenge) werden zwei Zahlen (Modell)

Trainingsmenge

	x	y	x*y	x^2	predicted	E^2
	2	2	4	4	2,3	0,09
	4	5	20	16	4,1	0,81
	6	5	30	36	5,9	0,81
	8	8	64	64	7,7	0,09
Summe	20	20	118	120		1,8
N	4					

Zähler	72
Nenner	80
w1 = m	0,90
w0 = n	0,50

Modell



„Lernen"

$$w_0 = \frac{1}{N}\sum y_i - \frac{w_1}{N}\sum x_i$$

$$w_1 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

Übung

Numerische Vorhersage - Beispiel

Wie viele Prüfungsbögen sollen zur Nachprüfung gedruckt werden?

Daten:

Prüfungsteilnahme bei den Pflichtmodulen im Nachprüfungszeitraum				
	Fach	Angemeldet	Teilgenommen	Semester
1. Semester Bachelor	Informatik und Logik	79	16	1
	Algorithmen und Datenstrukturen	75	14	1
	Technische Informatik und Medientechni	62	17	1
	Programmierung I	88	26	1
	Mathematik I	104	43	1
	Grundlagen der Medizininformatik I	20	3	1
	Grundlagen der Medizin I	21	8	1
3. Semester Bachelor	Programmieren III	60	42	3
	Betriebssysteme/Rechnernetze	23	15	3
	Grundlagen der Sicherheit	32	14	3
	Datenbanken I	30	13	3
	Mathematik III	47	29	3
	Grundlagen der Medizin III	7	2	3
	Medizinische Statistik und Biometrie	11	6	3
5. Semester Bachelor	Informatik und Gesellschaft	3	0	5
	BWL	18	12	5
	Computerunterstütze Medizin II	2	1	5
	BWL und Qualitätsmanagement	1	1	5

Modell:

$$Teilgenommen = 0.00398\, Angemeldet^2\, Semester$$

Numerische Vorhersage - Beispiel

- Es gäbe einen Benchmark für Computer, der nach langer Rechenzeit einen „Performancewert“ ausgibt.
- **Sagen Sie aus den Konfigurationsdaten eines PCs (cycle time [ns], main memory [KB, min, max], cache size [KB, min, max], channels) den Performancewert voraus.**

Lineares Regressionsmodell p =

$$
\begin{aligned}
& 0.0661 * MYCT + \\
& 0.0142 * MMIN + \\
& 0.0066 * MMAX + \\
& 0.4871 * CACH + \\
& 1.1868 * CHMAX + \\
& -66.5968
\end{aligned}
$$

<u>Weiteres Beispiel:</u>

Gegeben sind Studentendatensätze mit Notenprofil. Erstellen Sie ein Modell zur Vorhersage der Note der Abschlussarbeit.

Möglich?

3. Aufgabe: Clustering

(auch Segmentierung, Gruppenbildung, Clusteranalyse)

Beschreiben Vorhersagen

Aufspaltung unklassifizierter Daten in interessante und sinnvolle Teilmengen / Gruppen

Datensätze Innerhalb eines Clusters möglichst ähnlich, zwischen den Clustern möglichst unähnlich

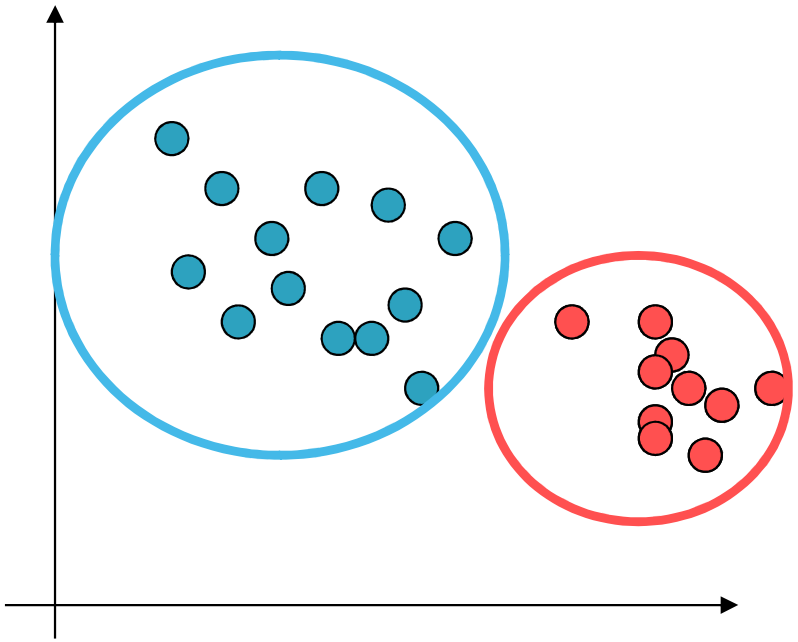
- Modelle:
  - Cluster-Repräsentanten: k-means, EM (Expectation–Maximization)
  - Selbstorganisierende Karten
  - Hierarchisches Clustering
  - …
- Andere Begriffe für Cluster: Cluster == Gruppen == Klassen == Teilmengen == Zusammenhangskomponenten == Nachbarschaften == Subgruppen == Segmente == Komponenten == …

Clustering - Beispiel

Kundensegmentierung

„Die Anwendung auf das Kundenverhalten einer Mobilfunk Gesellschaft ermöglichte eine Klassifizierung der Kunden in verschiedene Gruppen. Hierin konnte neues Wissen über das Verhalten von Kundensegmenten gewonnen werden. Von besonderem Interesse waren die Kunden, die mit hoher Wahrscheinlichkeit bald den Vertrag kündigen. „

[Ult04]



k-means

- Extrem einfach
- K-means einer der wichtigsten Algorithmen im Data Mining, siehe:

Knowl Inf Syst (2008) 14:1–37
DOI 10.1007/s10115-007-0114-2

SURVEY PAPER

Top 10 algorithms in data mining

Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg

k-means

- **Start**: wähle zufällig k Clusterzentren

- **Wiederhole**:
  - Ordne die Datenpunkte dem nächsten Clusterzentrum zu
  - Berechne neue Clusterzentren als Mittelwert aller zugeordneten Datenpunkte

- **Bis** sich die Zuordnung nicht mehr ändert.

Clustering k-means


Knowl Inf Syst (2008) 14:1–37
DOI 10.1007/s10115-007-0114-2
SURVEY PAPER
Top 10 algorithms in data mining
Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg

The k-means algorithm is a simple iterative method to partition a given specified number of clusters, k. This algorithm has been discovered by several ... across different disciplines, most notably Lloyd (1957, 1982) [53], Forgey (1965), Friedman and Rubin (1967), and McQueen (1967). A detailed history of k-means alongwith descriptions of several variations are given in [43]. Gray and Neuhoff [34] provide a nice historical background for k-means placed in the larger context of hill-climbing algorithms.

The algorithm operates on a set of d-dimensional vectors, $D = \{\mathbf{x_i} \mid i = 1, \ldots, N\}$, where $\mathbf{x_i} \in \Re^d$ denotes the ith data point. The algorithm is initialized by picking k points in $\Re^d$ as the initial k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence:

Step 1: Data Assignment. Each data point is assigned to its *closest* centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

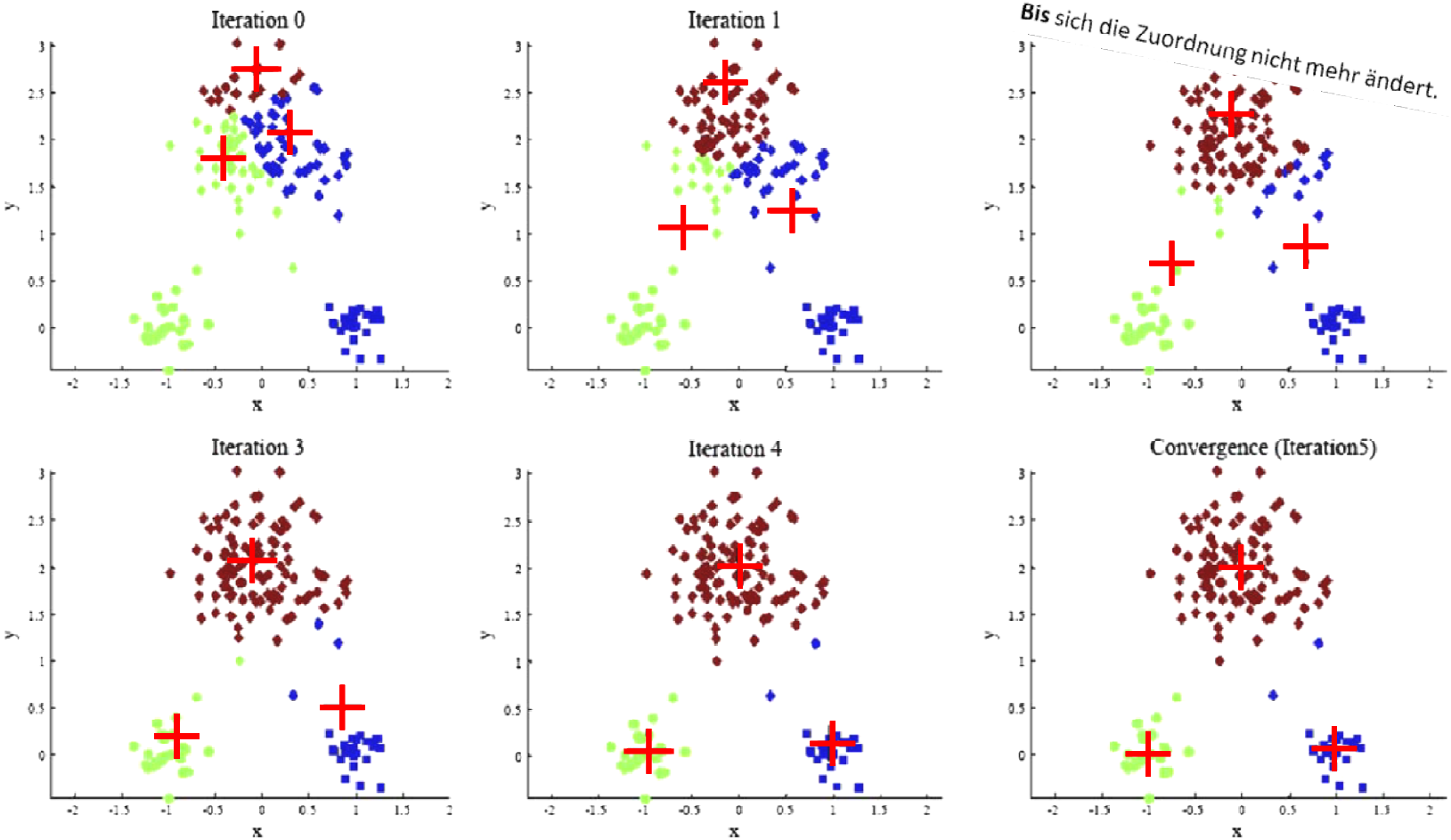
The algorithm converges when the assignments (and hence the $\mathbf{c_j}$ values) no longer change.

k-means Beispiel

Start: wähle zufällig k Clusterzentren

Wiederhole:
- Ordne die Datenpunkte dem nächsten Clusterzentrum zu
- Berechne neue Clusterzentren als Mittelwert aller zugeordneten Datenpunkte

Bis sich die Zuordnung nicht mehr ändert.



Übung

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. 2007. ***Top 10 algorithms in data mining****. Knowl. Inf. Syst. 14, 1 (December 2007)*

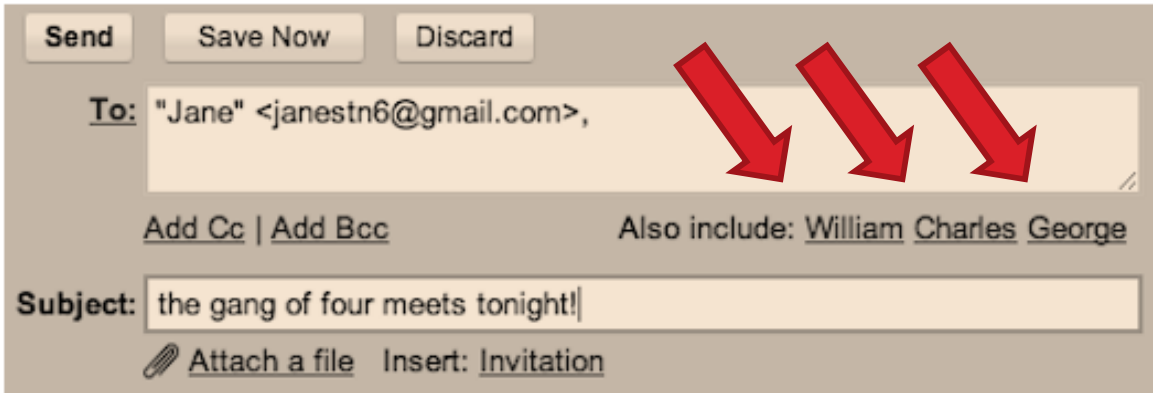
Clustering - Anwendungsbeispiel

- Homogene Gruppen erkennen
- Kunden, die zwischen Gruppen wandern
- Betrug: untypische Transaktionen erkennen, ausserhalb bekannter Cluster!

Beispiel Gmail:

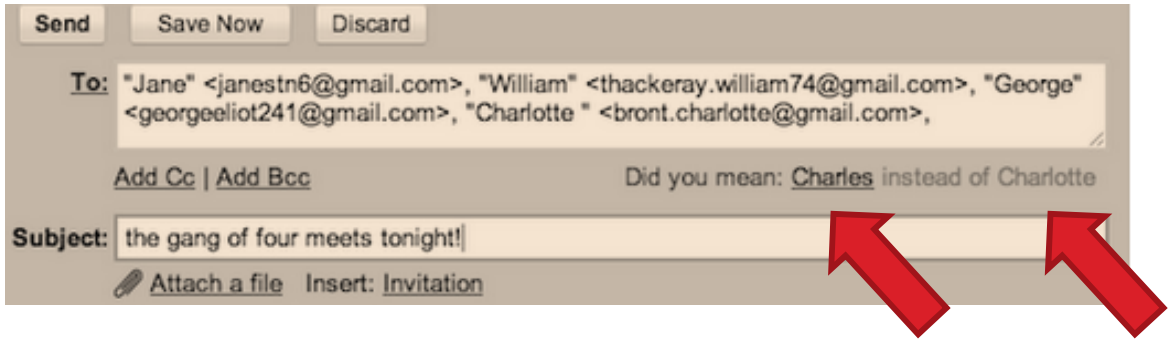
“Don't forget Bob”

(denn Bob ist auch im Cluster der Addressaten)



“Got the wrong Bob?”

(denn Bob ist nicht im Cluster der Addressaten)



[http://gmailblog.blogspot.com/2011/04/dont-forget-bob-and-got-wrong-bob.html]

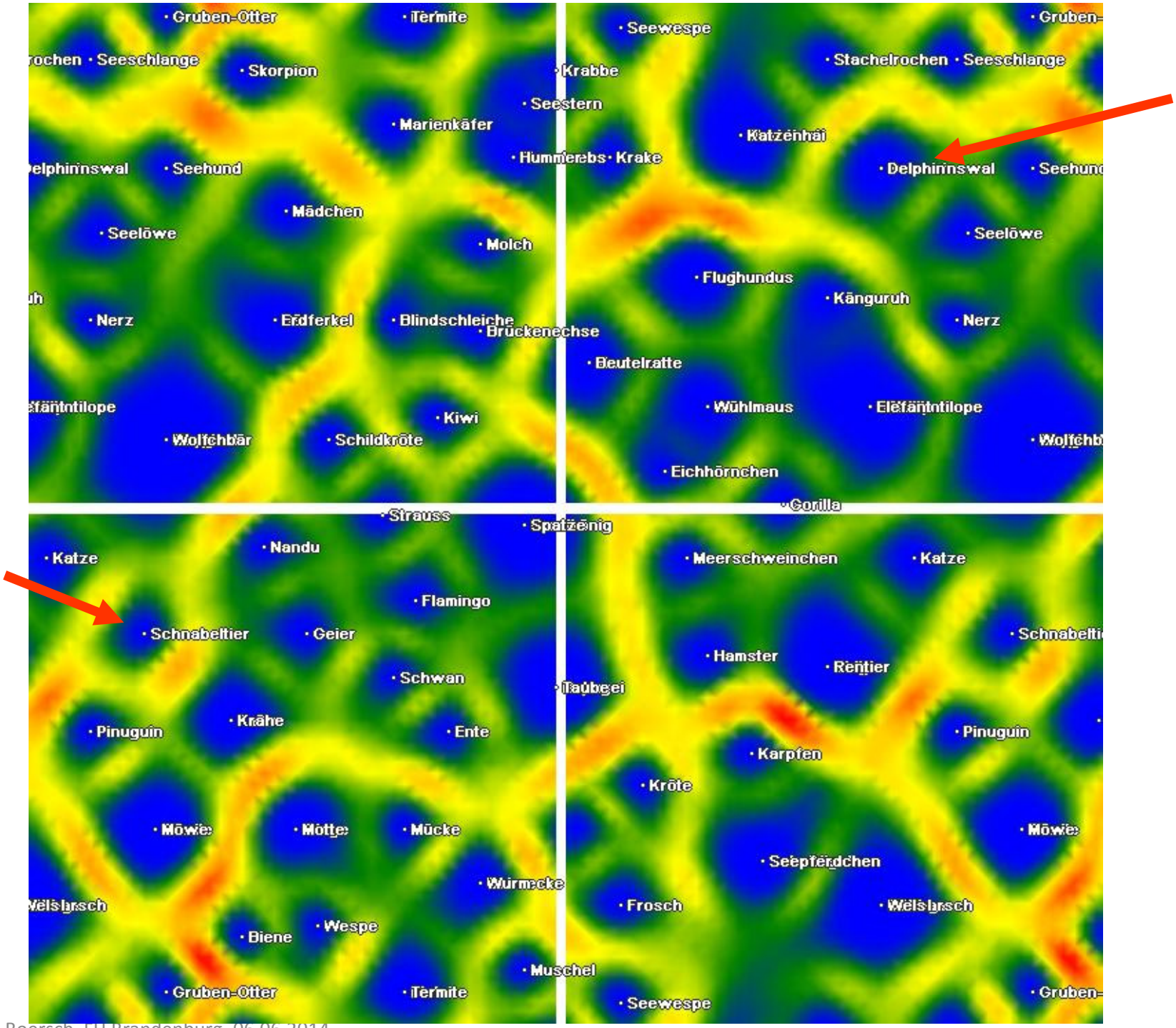
Clustering - Selbstorganisierende Karten

- gegeben: Tiere durch Eigenschaften beschrieben
- gesucht: Anordnung auf einer 2D-Karte so, dass ähnliche Tiere benachbart sind

Diplomarbeit von **Benjamin Hoepner**:

Entwurf und Implementierung einer Applikation zur Visualisierung von Lernvorgängen bei Selbstorganisierenden Karten

- Sombrero: 80x90, torus, a=1; exp=1; r=50; exp=0.995; s=20000
- Freie Software: http://ots.fh-brandenburg.de/diplomarbeit-von-benjamin-hoepner.html


Gruben-Otter
Termite
rochen
Seeschlange
Skorpion
Krabbe
Seestern
Marienkäfer
elphinnswal
Seehund
Hummerebs
Krake
Mädchen
Seelöwe
Molch
uh
Nerz
Erdferkel
Blindschleiche
Brückenechse
efäntntilope
Kiwi
Wolfchbär
Schildkröte
Seewespe
Gruben-
Stachelrochen
Seeschlange
Katzenhai
Delphinnswal
Seehund
Seelöwe
Flughundus
Känguruh
Nerz
Beutelratte
Wühlmaus
Elefäntntilope
Wolfchb
Eichhörnchen
Gorilla
Strauss
Spatzenig
Katze
Nandu
Flamingo
Schnabeltier
Geier
Schwan
Taubegei
Krähe
Pinuguin
Ente
Möwe
Motte
Mücke
Wurmecke
Welsbrsch
Biene
Wespe
Muschel
Gruben-Otter
Termite
Meerschweinchen
Katze
Schnabelti
Hamster
Rentier
Pinuguin
Karpfen
Kröte
Möwe
Seepferdchen
Frosch
Welsbrsch
Seewespe
Gruben-

4. Aufgabe: Abhängigkeitsanalyse

- Finden von Abhängigkeiten zwischen Attributen
- Modelle:
  - **Assoziationsregeln** (Finden mit Apriori-Algorithmus)
- Assoziationsregeln können
  - jedes einzelne Attribut vorhersagen, nicht nur das Klassenattribut
  - mehrere Attributwerte gleichzeitig vorhersagen
- Beispiel Warenkorbanalyse:
  - Produkte, die häufig zusammen gekauft werden
  - „Wenn Cornflakes und Milch gekauft werden, dann auch oft Zahnpasta“
  - Kombinationen von Sonderausstattungen

Abhängigkeitsanalyse - Beispiel

Folie enthält einen Fehler: Support falsch definiert

Regel: Linke Seite → Rechte Seite

- **Support** einer Regel: Häufigkeit der linken Seite, Anwendbarkeit
- **Konfidenz** einer Regel: wie oft trifft die Regel bei erfüllter linker Seite zu (Prozent)

Warenkorbanalyse sucht Regeln mit **hohem Support und hoher Konfidenz**.

Nehmen wir an, folgende Regel wird gefunden: Windeln → Bier

Was können Sie schlussfolgern?

A) Es werden oft Windeln gekauft.

~~*B) Der Kauf von Windeln führt zum Kauf von Bier.*~~

C) Wenn Windeln gekauft werden, dann oft auch Bier.

~~*D) Wenn keine Windeln gekauft werden, dann auch kein Bier.*~~

Übung

- Vorsicht bei B: **Korrelationen sind keine Kausalzusammenhänge**

Abhängigkeitsanalyse - Beispiel

Regel: Linke Seite → Rechte Seite

- **Support** einer Regel: Wie oft treten linke und rechte Seite gemeinsam auf
- **Konfidenz** einer Regel: wie oft trifft die Regel bei erfüllter linker Seite zu (Prozent)

Warenkorbanalyse sucht Regeln mit **hohem Support und hoher Konfidenz**.

Nehmen wir an, folgende Regel wird gefunden: Windeln → Bier

Was können Sie schlussfolgern?

A) Es werden oft Windeln gekauft.

~~*B) Der Kauf von Windeln führt zum Kauf von Bier.*~~

C) Wenn Windeln gekauft werden, dann oft auch Bier.

~~*D) Wenn keine Windeln gekauft werden, dann auch kein Bier.*~~

Übung

- Vorsicht bei B: **Korrelationen sind keine Kausalzusammenhänge**

Beispiel Support und Konfidenz

Beispiel aus:

Bollinger T.: *Assoziationsregeln - Analyse eines Data Mining Verfahrens*, in Informatik Spektrum 5/96, S. 257 ff.

Einkaufstransaktion	gekaufte Artikel (Item)
t_1	Saft, Cola, Bier
t_2	Saft, Cola, Wein
t_3	Saft, Wasser
t_4	Cola, Bier, Saft
t_5	Saft, Cola, Bier, Wein
t_6	Wasser

Artikel	Transaktionen, in denen der Artikel vorkommt
Saft	t_1, t_2, t_3, t_4, t_5
Cola	t_1, t_2, t_4, t_5
Bier	t_1, t_4, t_5
Wein	t_2, t_5
Wasser	t_3, t_6

$$\text{Support}(Saft \rightarrow Cola) = \frac{|\text{t mit Saft und Cola}|}{|\text{alle}|} = \frac{4}{6} \approx 67\%$$

$$\text{Konfidenz}(Saft \rightarrow Cola) = \frac{|\text{t mit Saft und Cola}|}{|\text{t mit Saft}|} = \frac{4}{5} = 80\%$$

Regeln mit Support $\geq$ 50 %	erfüllende Transaktionen	Support	Konfidenz
$Saft \rightarrow Cola$	t_1, t_2, t_4, t_5	66 %	80 %
$Cola \rightarrow Saft$	t_1, t_2, t_4, t_5	66 %	100 %
$Cola \rightarrow Bier$	t_1, t_4, t_5	50 %	75 %
$Bier \rightarrow Cola$	t_1, t_4, t_5	50 %	100 %

Korrelation ≠ Kausalität

Vorsicht: **Korrelationen sind keine Kausalzusammenhänge**

- (A) Benzinlampe leuchtet → (B) Auto bleibt stehen
- (A) Gelbe Finger → (B) Lungenkrebs
- (A) hohe Storchendichte → (B) hohe Geburtenrate

Kausalität hinter einer Korrelation A→ B könnte sein:

1. A => B, also A ist Ursache von B, oder
2. B => A oder
3. eine gemeinsame Ursache C: C => A und C => B

Korrelationen geeignet zur Vorhersage (**predict**):
Bei gelben Fingern ist mit Lungenkrebs zu rechnen.

Kausale Beziehungen geeignet für die gezielte Beeinflussung (**manipulate**):
Das Schrubben gelber Finger senkt das Krebsrisiko nicht.

Dennoch ist es eine gute Idee, für mehr Geburten die Storchendichte zu erhöhen.

5. Aufgabe: Abweichungsanalyse

- Ausreißer: untypische Merkmalsausprägungen, bspw. mehr als 3 Standardabweichungen vom Mittelwert (3-Sigma-Regel, 4, 5 ..) entfernt

Analyse: Lässt sich die Ursache finden?

  - Falsche oder alte Annahmen über den Datengenerierungsprozess (z.B. Hat sich der Prozess unbemerkt geändert?)
  - Messfehler, Rauschen?

  -> erkennen, entfernen, weniger beachten oder reparieren

- sonst (Ausreißer ist zwar überraschend, aber **korrekt**)
  - Fehlen wichtige Attribute oder Einflußgrößen?
  - Ausreißer erkennen und extra trainieren

Des einen Ausreißer ist des anderen Datenpunkt.

[Text-Mining]

Eher ein **Anwendungsgebiet**, weniger eine Aufgabenart

- **Klassifikation** von Texten (Thema, Kategorie, Stimmungslage, bester Ansprechpartner für …, Spam-Filter)
- Extrahieren von Informationen (Wer was wann wo, Entity Mapping)
- Strukturieren von Textmengen (**Clustern** für Bibliothekskataloge)

Modelle

- Wie Klassifikation und Clustering!!
- Besonderheit ist die **Bestimmung von Merkmalen** (engl. feature extraction) aus Texten

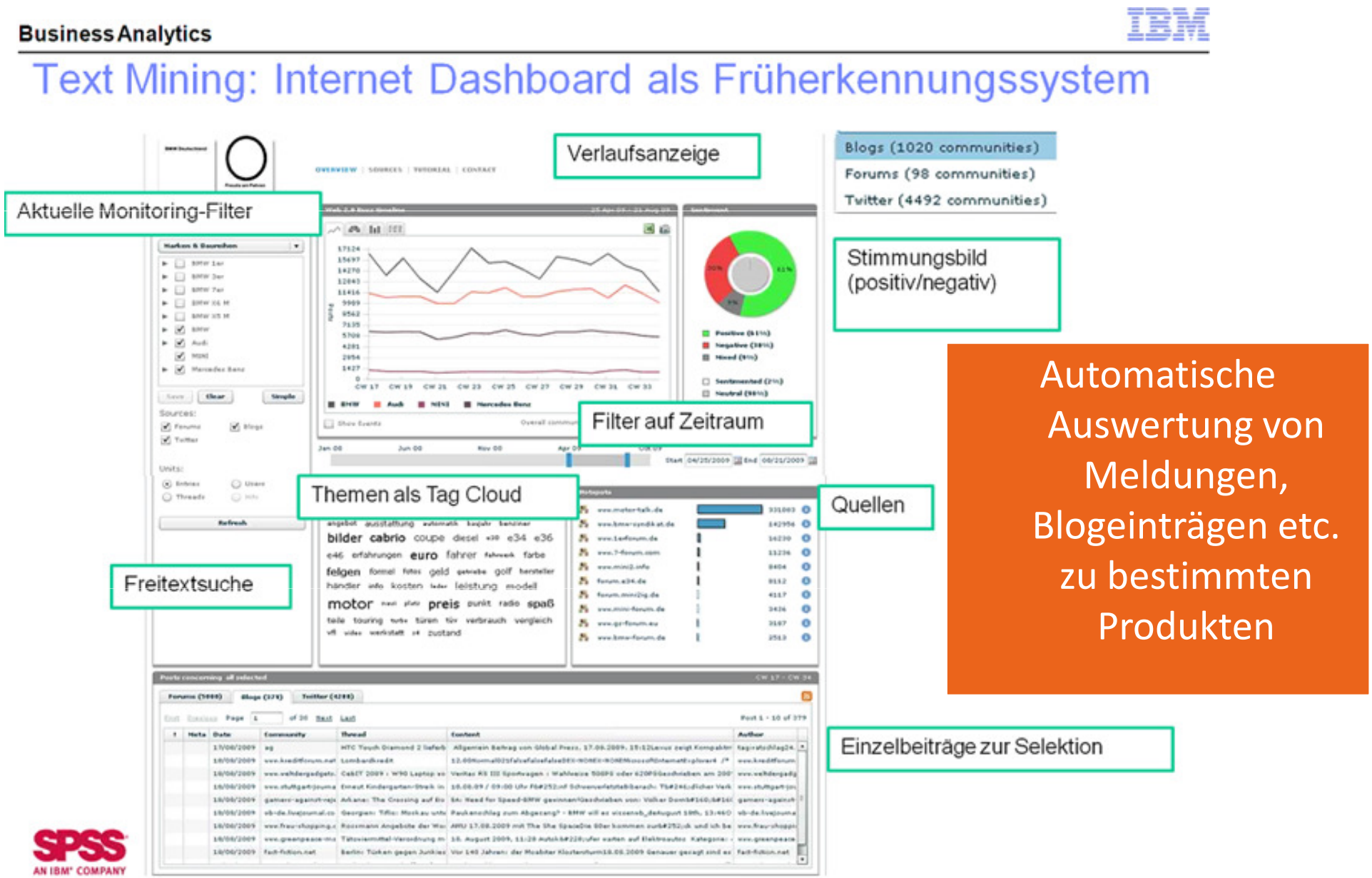
Beispiele

- Patentanalyse
- **Welche Themen** werden im Netz in unserem Kontext diskutiert?

  .. sind positiv oder negativ besetzt,

  .. werden von abwanderungsgefährdeten Kunden im Kundencenter angesprochen?

Beispiel Text-Mining



Quelle: Webinar „Data Mining in der Fertigung und Qualitätsoptimierung", Anja Burkhardt, Mai 2010

Sie kennen jetzt die Aufgabenstellungen des Data Mining

- ☑ Klassifikation
- ☑ Numerische Vorhersage
- ☑ Abhängigkeitsanalyse
- ☑ Clustering
- ☑ Abweichungsanalyse

- ☑ [Text-Mining]

- ☑ [lineare Regression]
- ☑ Support, Konfidenz einer Regel
- ☑ k-means

Next

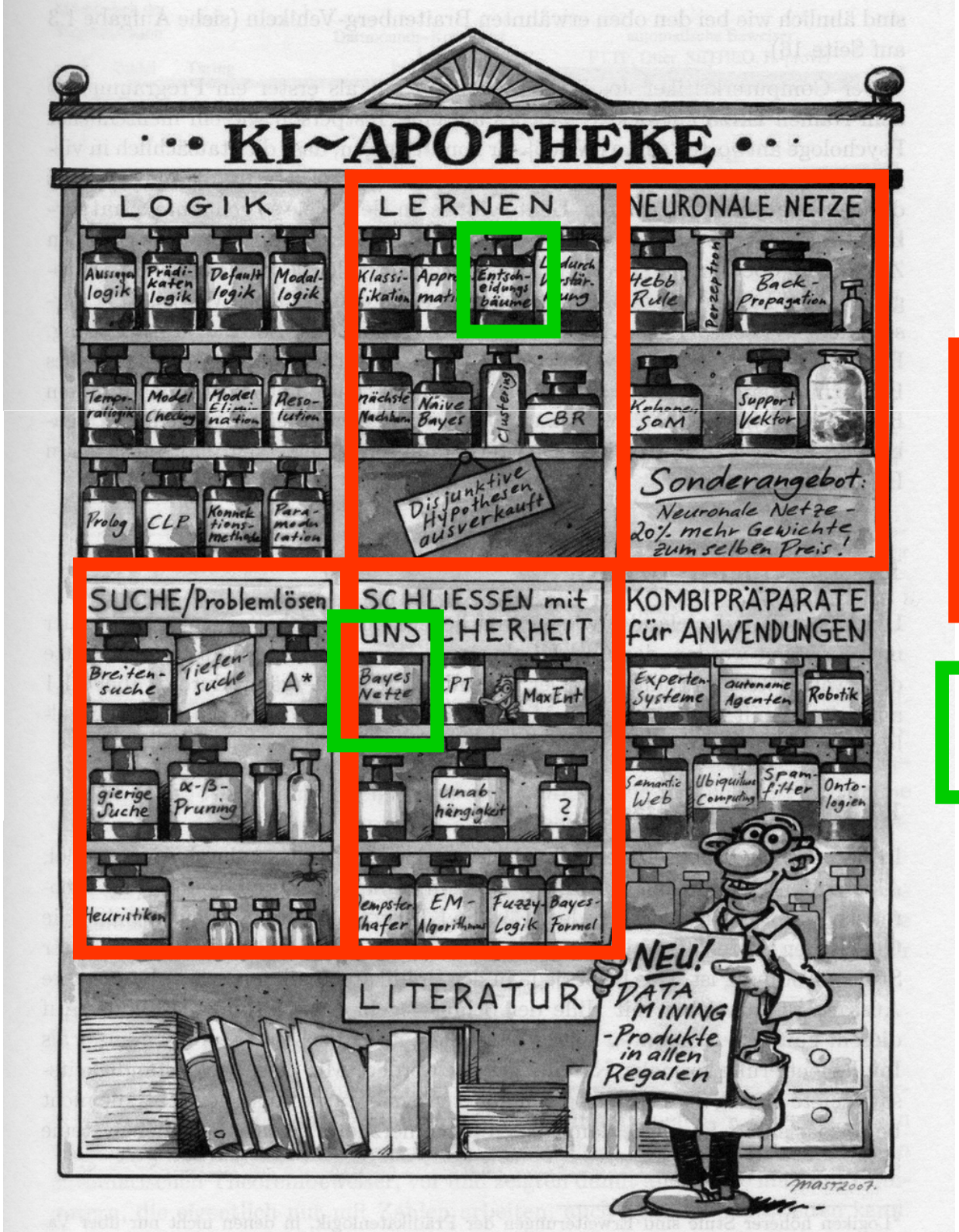
- DM und KDD, Phasen
- Aufgabenstellungen des DM

- **Wissensrepräsentation**
  - KI-Apotheke
  - Transparenz
  - KD-Nuggets

- Entscheidungsbäume I - Repräsentation
- Entscheidungsbäume II - Lernen
- Entscheidungsbäume III - Praktisch
- Performance von Klassifikatoren
- Ethik

Anwendbarkeit beim Data Mining

Abb. aus W. Ertel, Grundkurs KI, 2008



Anwendbarkeit beim Data Mining

in der KI-Vorlesung

Abb. aus W. Ertel, Grundkurs KI, 2008

Transparenz (Verständlichkeit der Modelle)

Eigenschaften von **Wissensrepräsentationen**: Vollständigkeit, Abstraktion, Ökonomie, Freiheit von Redundanz, Transparenz, Erweiterbarkeit, Erlernbarkeit ...

Transparenz ist beim Data Mining von besonderer Bedeutung

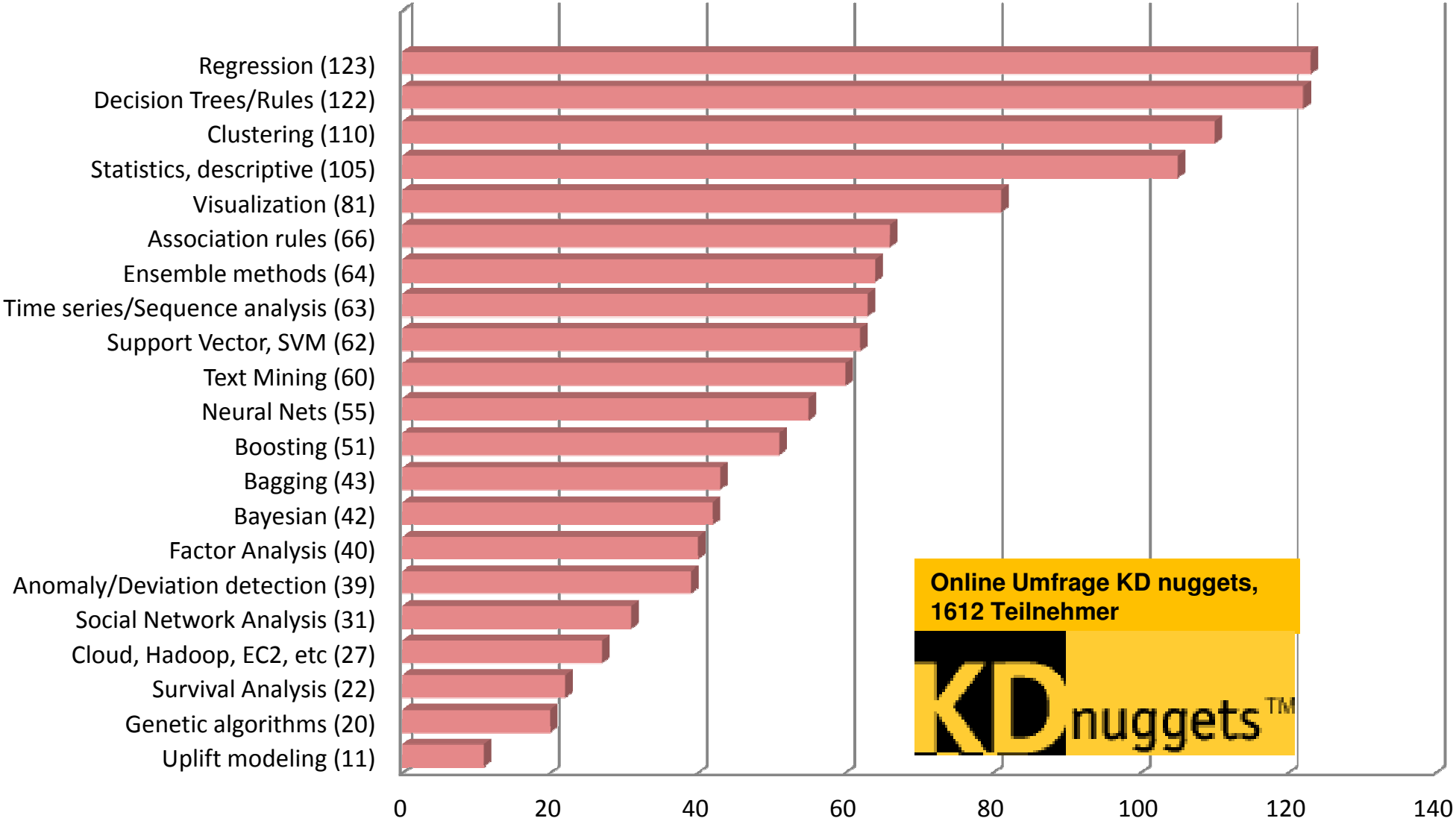
„Die Erfahrung hat gezeigt, dass in vielen Anwendungen des maschinellen Lernens für das Data Mining die expliziten Wissensstrukturen, die strukturierten Beschreibungen, mindestens ebenso wichtig und nicht selten wichtiger sind als die Fähigkeit, eine gute Leistung für neue Beispiele zu erbringen.“

Ian Witten, Eibe Frank in WF01

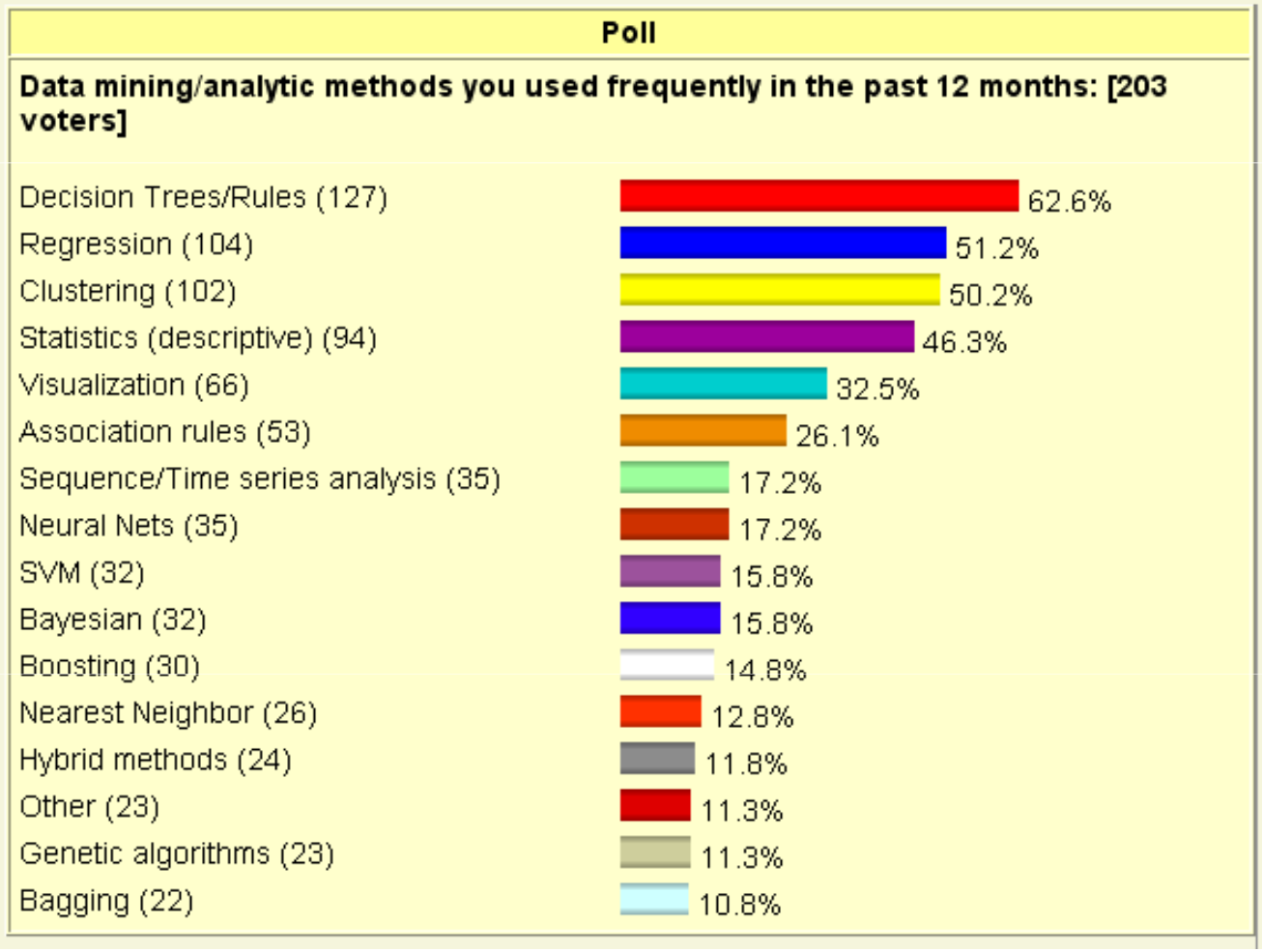
Beschreiben

Vorhersagen

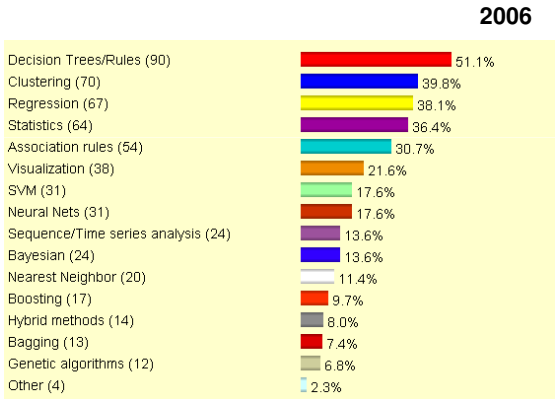
Which methods/algorithms did you use for data analysis in 2011?



Häufig genutzte Methoden



Data mining/ analytic methods you used frequently in the last year
(Umfrage KD-Nuggets März 2007)



[http://www.kdnuggets.com/polls/index.html]

Next

- DM und KDD, Phasen
- Aufgabenstellungen des DM
- Wissensrepräsentation

- **Entscheidungsbäume I – Repräsentation**
  - Beispiel als ARFF
  - Klassifikation
  - Knoten
  - **Zerlegung des Merkmalsraumes**
  - Multivariate Bäume

- Entscheidungsbäume II - Lernen
- Entscheidungsbäume III - Praktisch
- Performance von Klassifikatoren
- Ethik

Ein Spiel an frischer Luft

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

ARFF-Datenformat (Weka, RapidMiner)

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

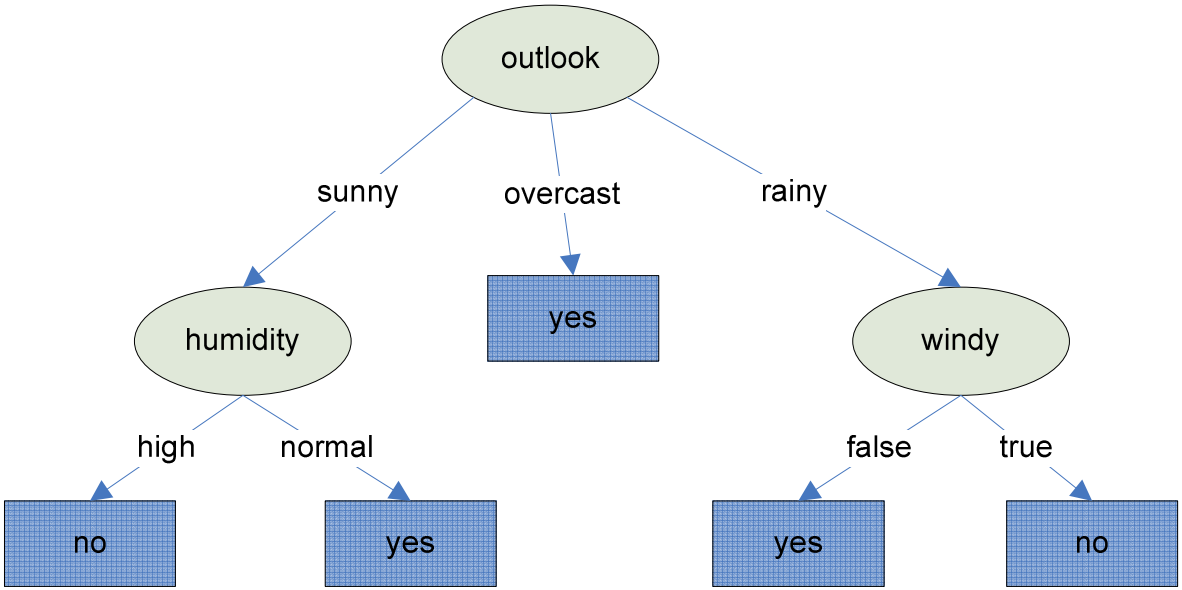
Fünf nominale Attribute

Skalenniveau von Attributen:

- nominale (Aufzählung, ähnlich enum-Typ)
- ordinale (wie enum aber mit Reihenfolge)
- numerische (einfache Zahlen)

Ein Entscheidungsbaum

- Jeder Knoten testet ein Attribut
- Jeder Zweig repräsentiert einen Attributwert
- Jeder Blattknoten weist eine Klassifikation zu

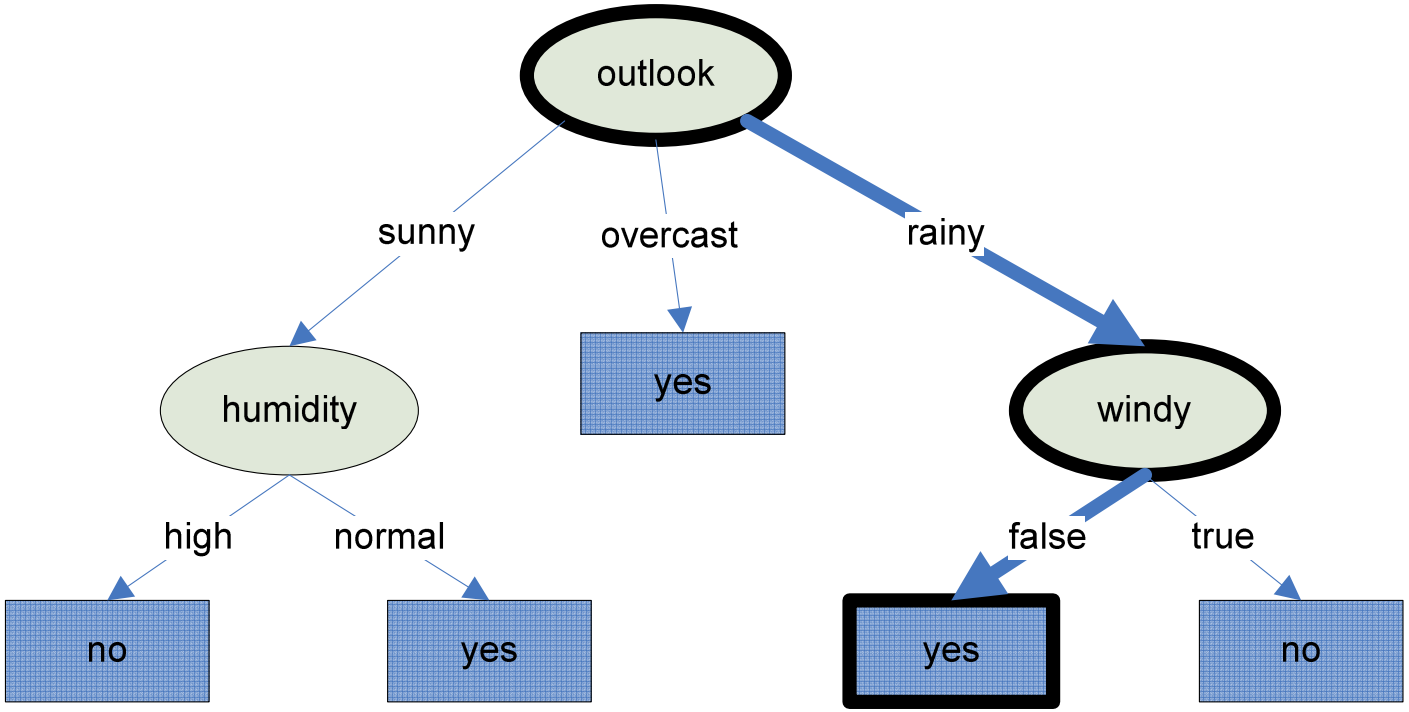


- Verschiedene Bäume möglich – Welcher ist der beste (einer der besten)? Wie finden wir ihn? Wie verwenden wir einen Entscheidungsbaum?

Wie wird klassifiziert?

Welche Klasse hat der Datensatz Nr. 5?

{Outlook=rainy, temperature=cool, humidity=normal, windy=false}



Testfunktionen in den Knoten

Die Art der Testfunktion ist entscheidend für die Transparenz der Bäume.

Häufig nur diese beiden Testfunktionen erlaubt:

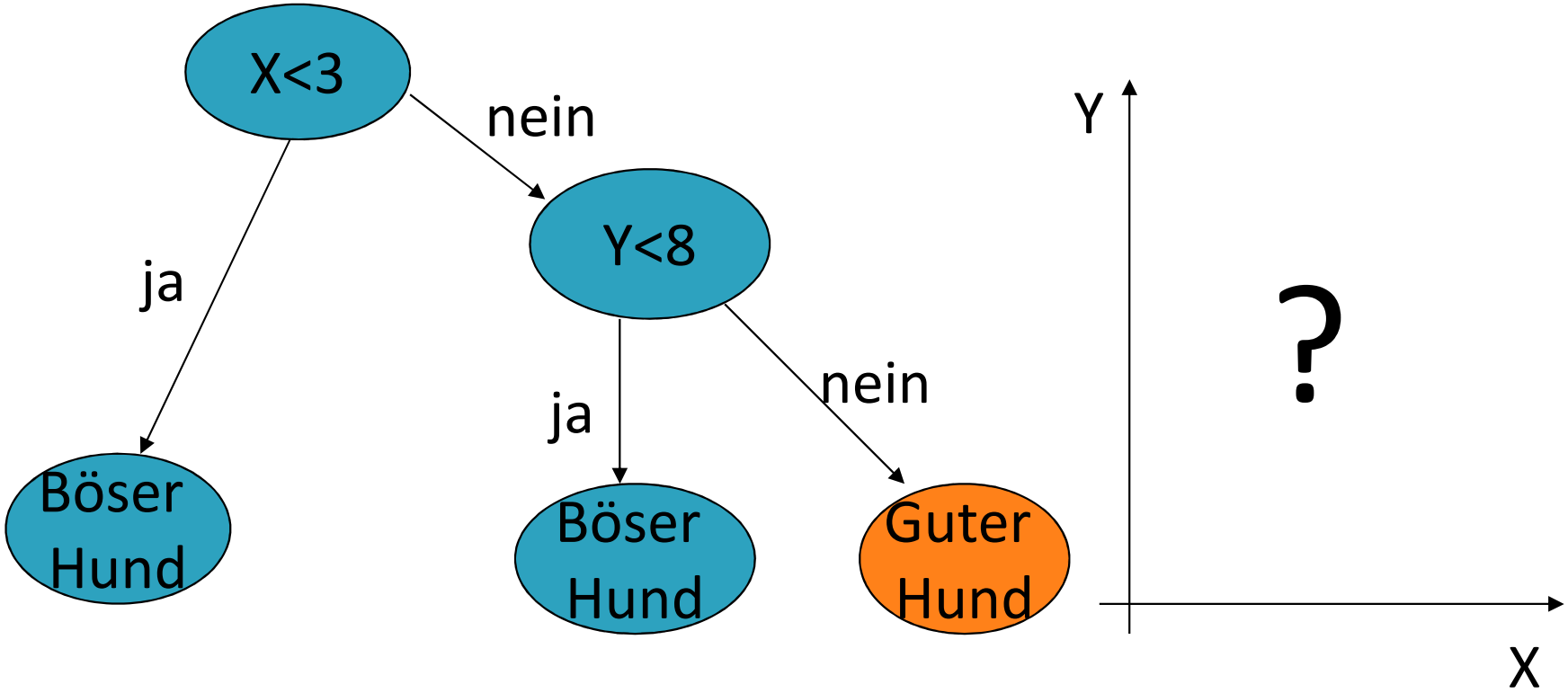
1. Nominales Attribut => Aufspaltung in alle Attributwerte
2. Numerisches Attribut verglichen mit Konstante => yes oder no

z. B. nicht erlaubt:

- temperature == humidity (mehrere Attribute)
- temperature == cool oder temperature = hot (mehrere Werte)

Zerlegung des Raumes

Zerlegung des Merkmalsraumes in **achsenparallele Hyperrechtecke** durch numerische Attribute



Zerlegung des Raumes

Zerlegung des Merkmalsraumes in **achsenparallele Hyperrechtecke** durch numerische Attribute

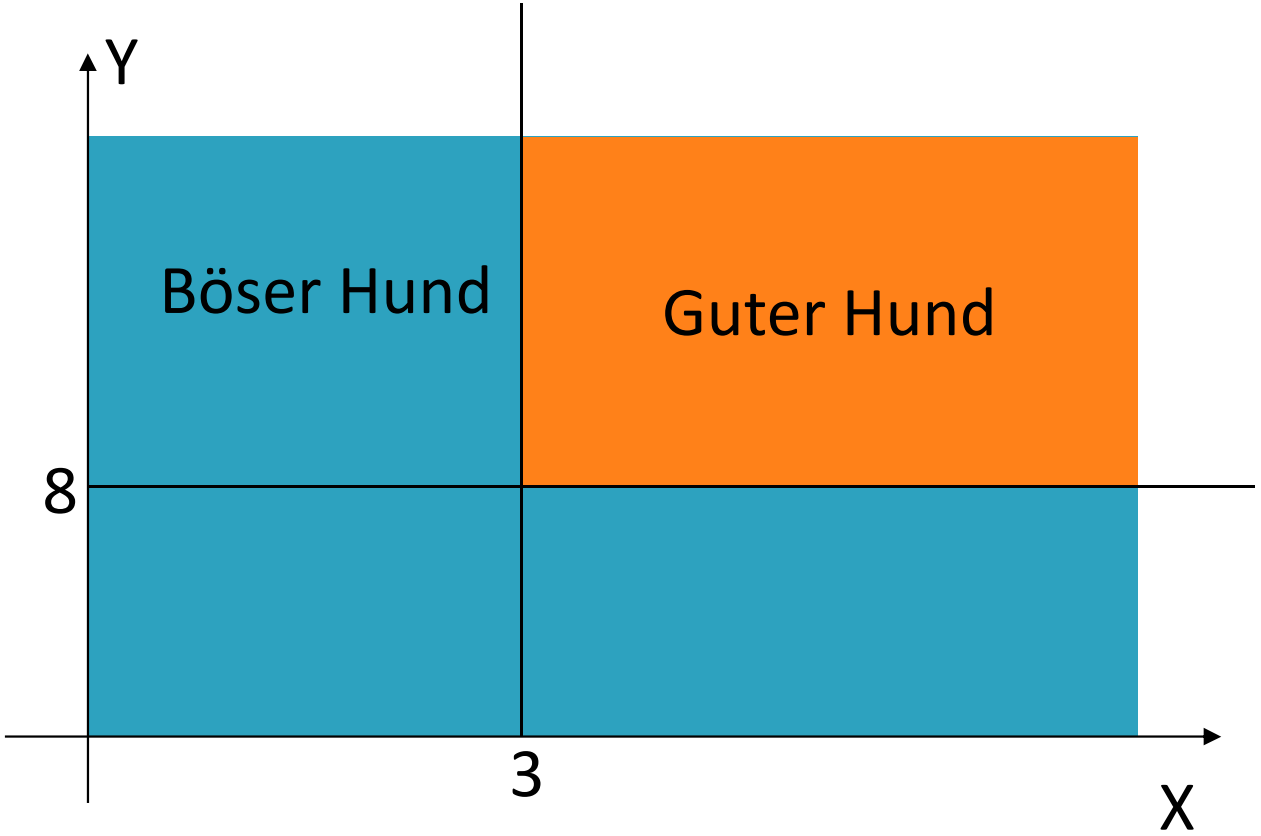
Textdarstellung des Baumes:

X < 3 : Böser Hund

X >=3 :

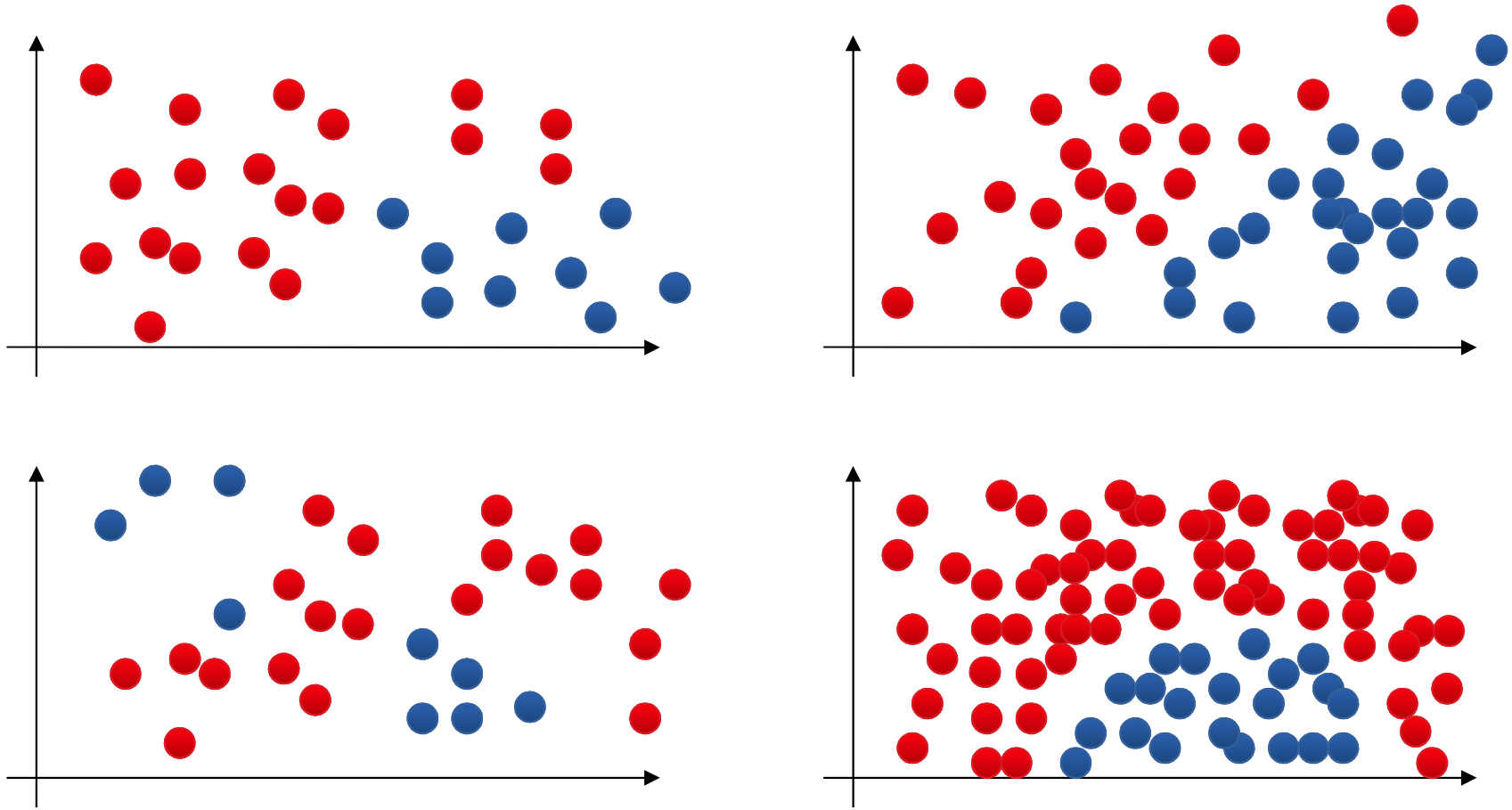
Y < 8 : Böser Hund

Y >= 8 : Guter Hund



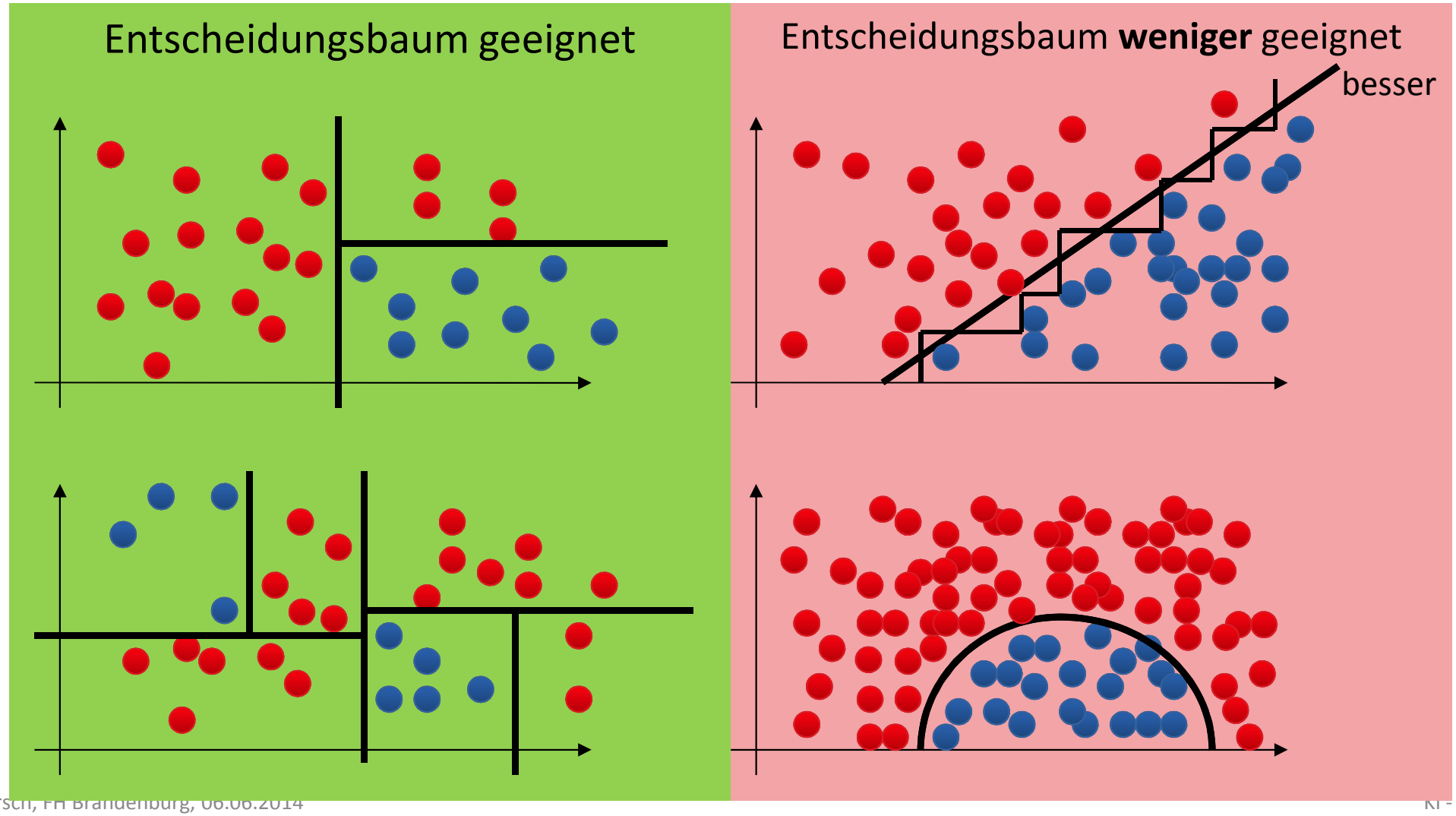
Zerlegung des Raumes

Zerlegung des Merkmalsraumes in **achsenparallele Hyperrechtecke** durch numerische Attribute – wann eignet sich ein Entscheidungsbaum?



Zerlegung des Raumes

Zerlegung des Merkmalsraumes in **achsenparallele Hyperrechtecke** durch numerische Attribute



Multivariate Entscheidungsbäume

- **Univariate** Bäume: jeder Knoten testet nur ein Attribut
- **Multivariate** Bäume: Knoten testen mehr als ein Attribut

Beispiele für multivariate Knoten:

- **Linearkombination von Attributen** als Knoten, z.B.

  *Gewicht + 2 * Größe < 70*

  => Schnitte im Merkmalsraum noch linear, aber nicht mehr achsenparallel

- **Nichtlineare Ausdrücke**, wie

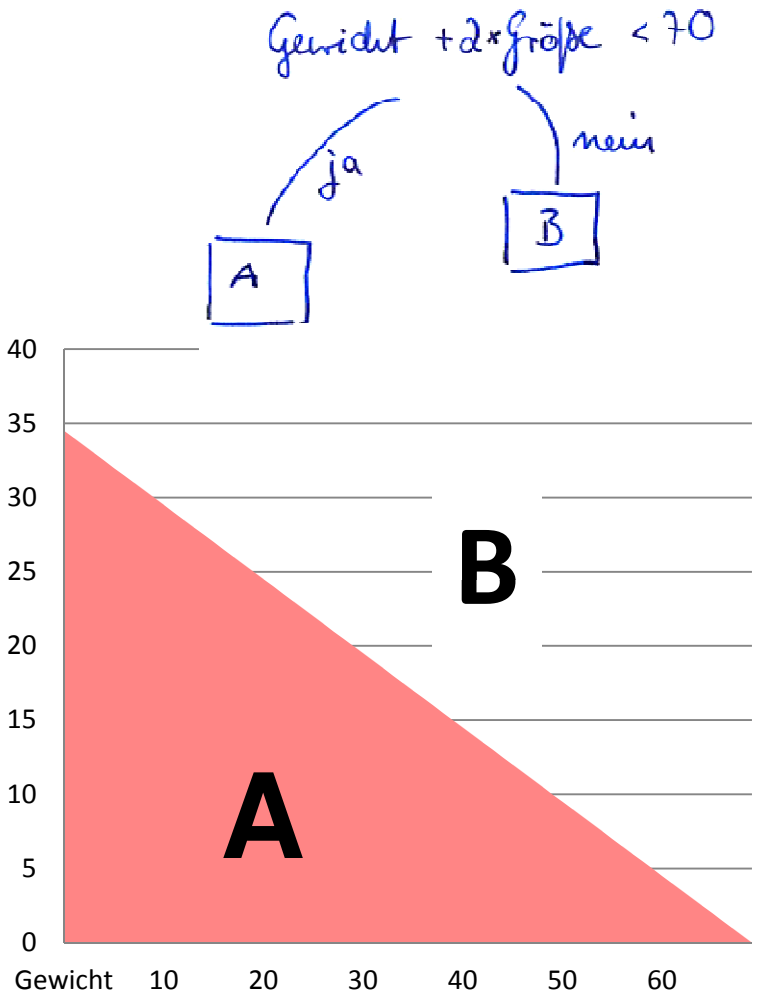
  *Gewicht / (Größe*Größe) < 25*

  => Schnitte im Merkmalsraum beliebig kurvig

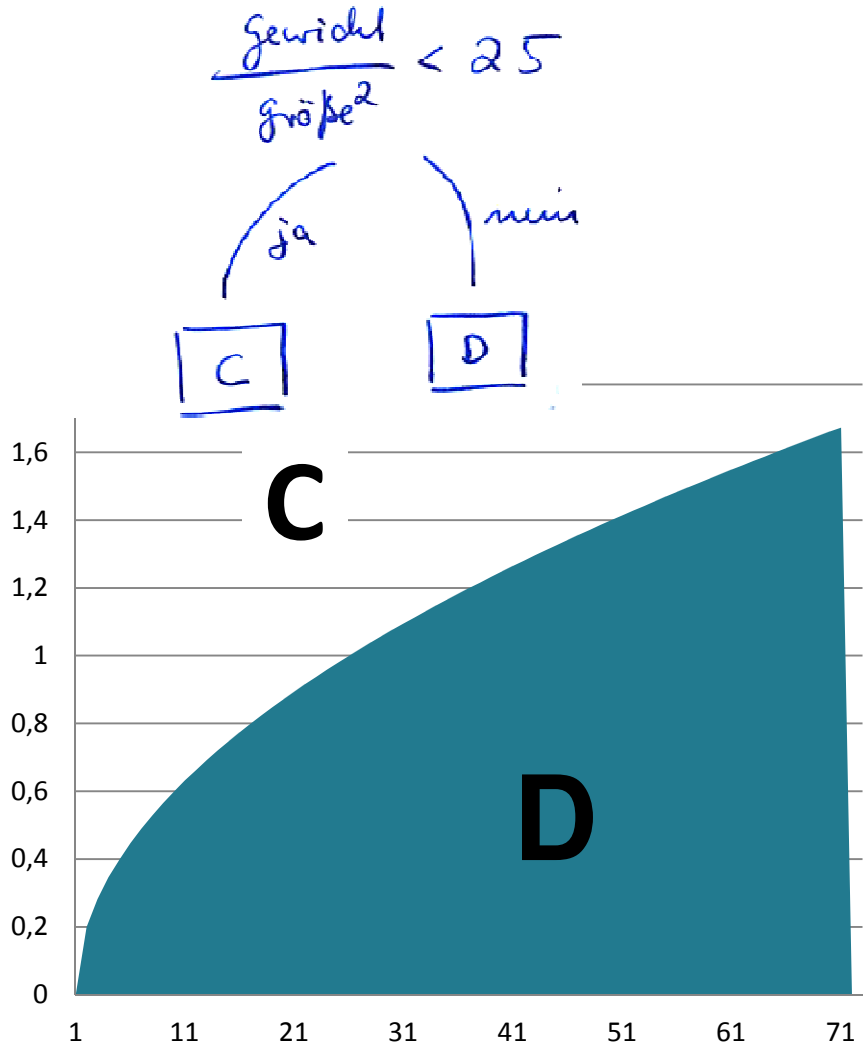
- **Vorteil**: oft genauer und kleiner,
- **Nachteil**: schwieriger zu bauen und zu interpretieren

Zerlegung des Merkmalsraumes bei **multivariaten** Bäumen

Linearkombination von Attributen:



Nichtlineare Ausdrücke:



Next

- DM und KDD, Phasen
- Aufgabenstellungen des DM
- Wissensrepräsentation
- Entscheidungsbäume I – Repräsentation

- **Entscheidungsbäume II – Lernen**
  - Lernalgorithmus ID3
  - Das beste Attribut
  - Entropie
  - Informationsgewinn Gain
  - Ergebnis des Beispiels

- Entscheidungsbäume III – Praktisch
- Performance von Klassifikatoren
- Ethik

HA Motive und Triebkräfte

Besorgen und verkaufen Sie personenbezogene Daten mit dem Online-Spiel „Data Dealer“ unter http://demo.datadealer.net/ bis Sieglinde Bayer-Wurz auftaucht.

1. Bei welchen Personen führen folgende Motive zur Datenweitergabe?
   - Schulden und Geldprobleme:
   - Unzufriedenheit mit der Arbeitstätigkeit :
   - Unzufriedenheit mit Lohn/Gehalt:
   - Erpressung:
   - Hacker:
   - Rache:

2. Welche beiden Datenquellen verkaufen die Daten, die für die Krankenversicherung am wertvollsten sind?

Lernen von Entscheidungsbäumen

Der beste Baum?

– Gültig, flach, geringe Knotenanzahl, gleichmäßige Datendichte, übersichtlich …
=> **„kleiner korrekter Baum“**

- Anzahl der möglichen Entscheidungsbäume gigantisch
- Durchprobieren aller Bäume undurchführbar

=> Ein Suchproblem! Ein Optimierungsproblem!

Lässt sich ein guter Baum schrittweise konstruieren?

Ja, Gierige (greedy) Strategie ähnlich Bergsteigen – bergauf losmarschieren und niemals umdrehen

Top-Down Induction of DecisionTrees (TDIDT)

TDIDT

Erzeuge Wurzelknoten *k*
Berechne TDIDT(*k*, Trainingsdaten)

Wirkt kompliziert, ist aber sehr einfach

TDIDT(*k,* Trainingsdaten):

1. Haben alle Beispiele die gleiche Klasse, so weise dem Knoten *k* diese Klasse zu (ein Blatt).
2. sonst:
   - Bestimme das **beste** Attribut *A* für eine Zerlegung der Trainingsdaten
   - Weise dem Knoten *k* den Test *A* zu
   - Bestimme Menge *T* aller möglichen Testergebnisse von *A*
   - Für alle Testergebnisse *t* aus *T*:
     - Erzeuge einen Nachfolgerknoten k_t zu *k*
     - Beschrifte die Kante von *k* nach k_t mit *t*
     - Setze Beispielmenge B_t = leere Menge
   - Für alle Beispiele *b* aus der Beispielmenge:
     - Wende den Test *A* auf *b* an und bestimme den Testausgang *t*
     - Füge das Beispiel *b* zur Menge B_t hinzu
   - Für jeden Nachfolgeknoten k_t von *k*:
     - Berechne TDIDT(k_t, B_t)

Attributwerte

Nachfolger erzeugen

Beispiele verteilen

Grundidee TDIDT

TDIDT reduziert die Suche nach dem besten Baum auf die Bestimmung des besten Attributes

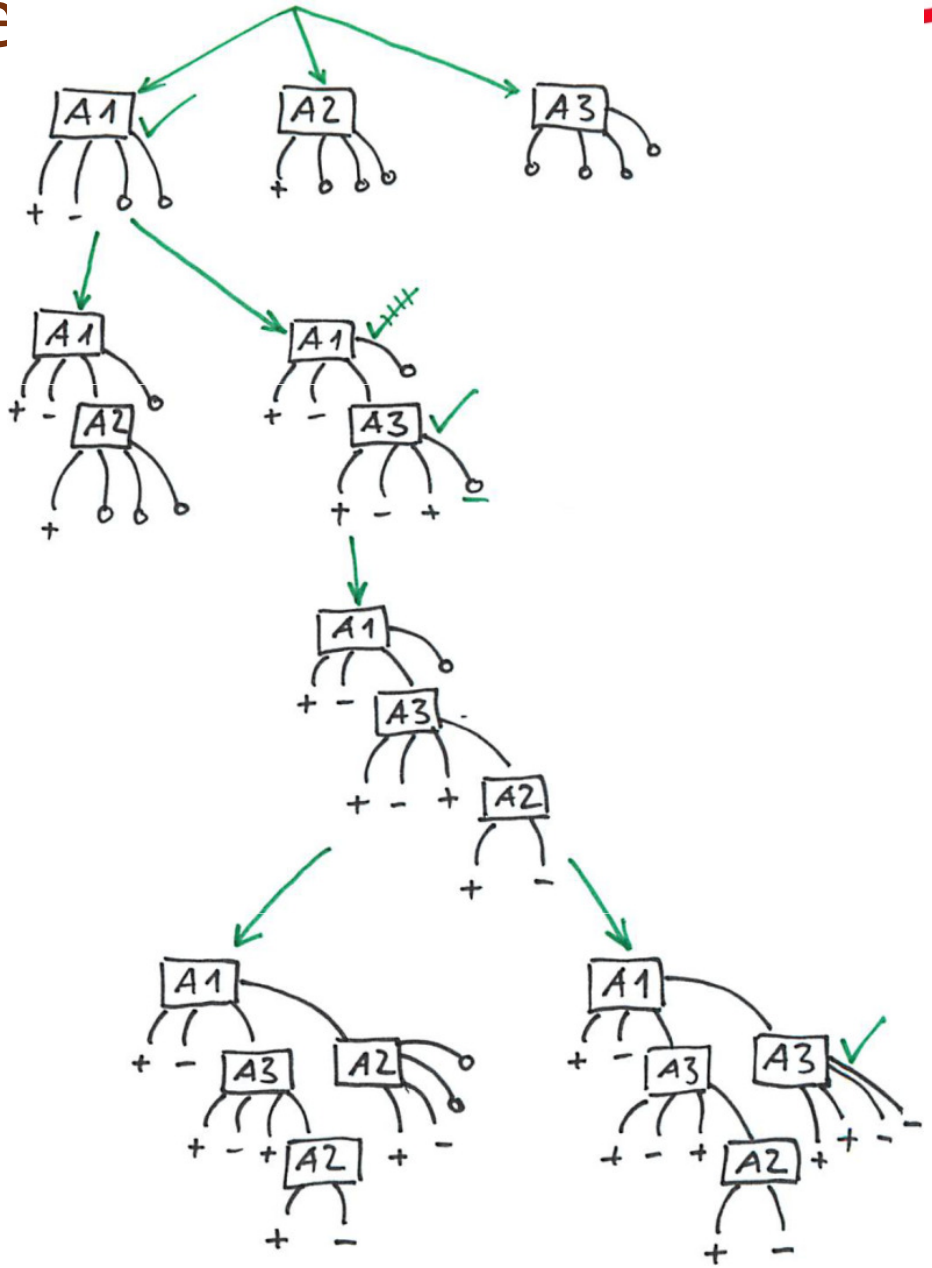
1. Reiner Knoten – Klasse, fertig für diesen Knoten
2. Wähle das **beste Attribut A** für den aktuellen Knoten.
3. Für jeden Wert von A erzeuge einen Nachfolgeknoten und markiere die Kante mit dem Wert.
4. Verteile die aktuelle Beispielmenge auf die Nachfolgeknoten, entsprechend den jeweiligen Werten von A.
5. Wende TDIDT auf alle neuen Nachfolgenoten an (Rekursion)

- Heuristische Suche (Schätzung des besten Attributes)
- Keine Optimalitätsgarantie
- Kein Backtracking

TDIDT: Verlauf der Suche

Beispiel (Tafel)

- 3 Attribute A1, A2, A3
- Jeweils 4 Werte
- 2 Klassen {+,-}



+ = reiner Knoten (Blatt)

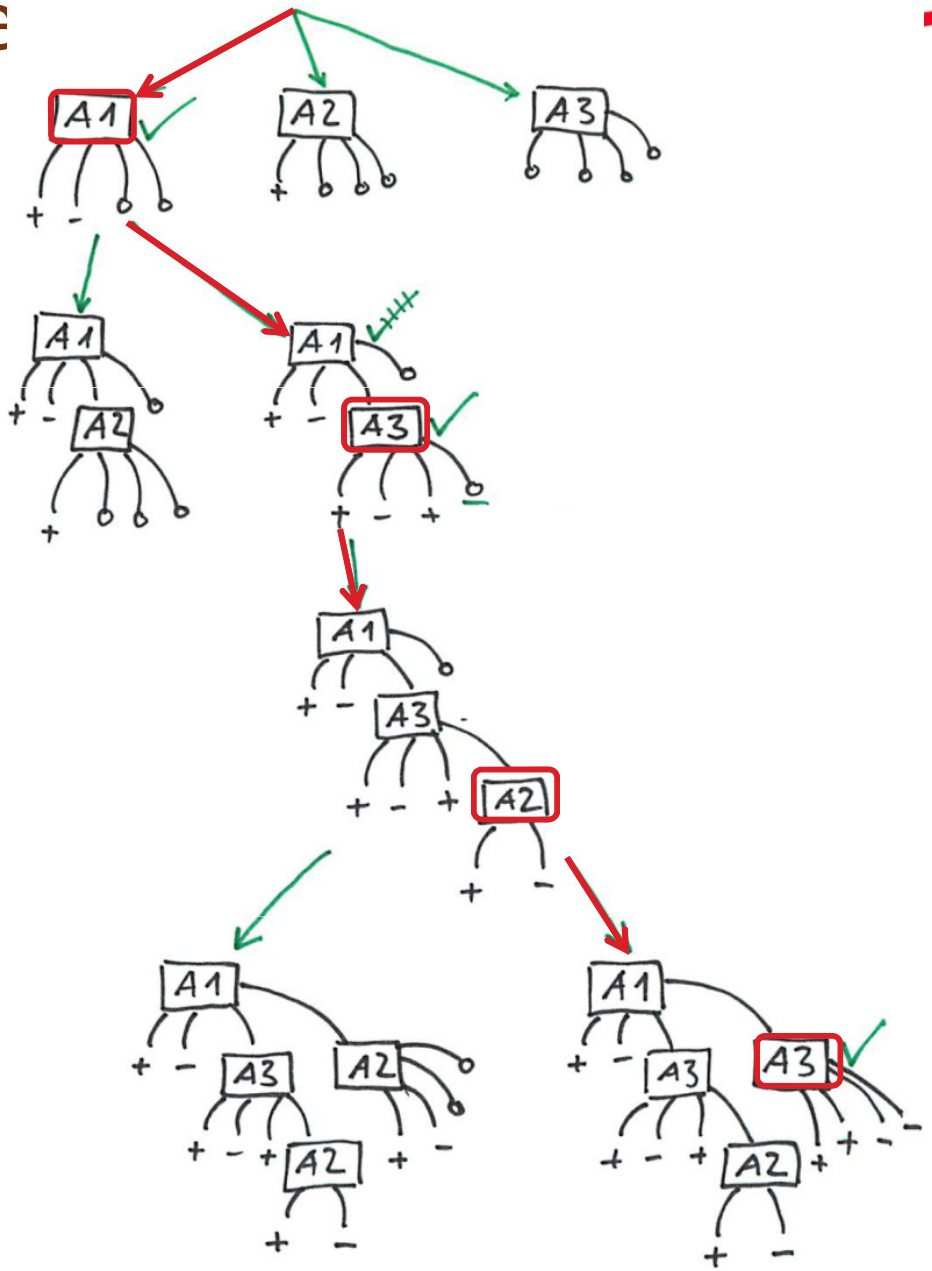
- = reiner Knoten (Blatt)

o = Knoten enthält Datensätze beider Klassen

TDIDT: Verlauf der Suche

Beispiel (Tafel)

- 3 Attribute A1, A2, A3
- Jeweils 4 Werte
- 2 Klassen {+,-}

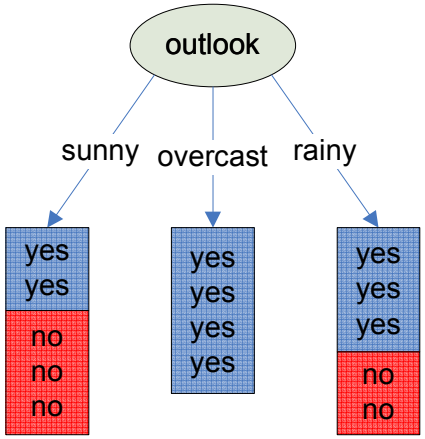


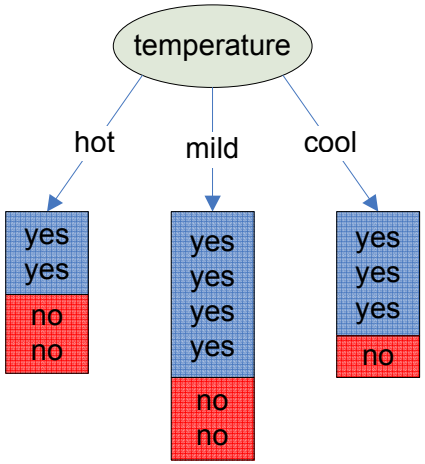
+ = reiner Knoten (Blatt)

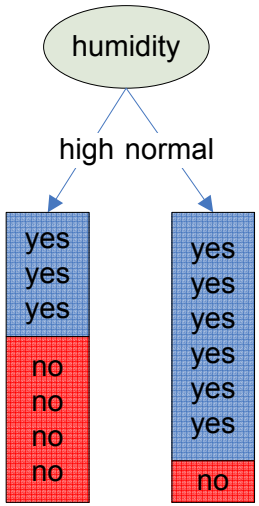
- = reiner Knoten (Blatt)

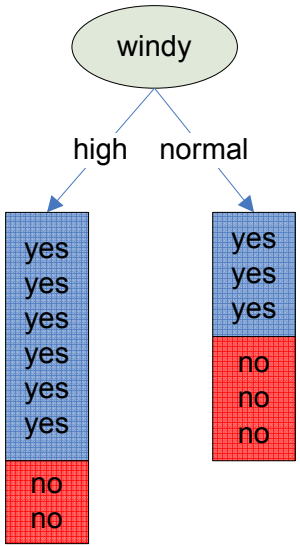
o = Knoten enthält Datensätze beider Klassen

Das beste Attribut?









Welches ist das beste Attribut?

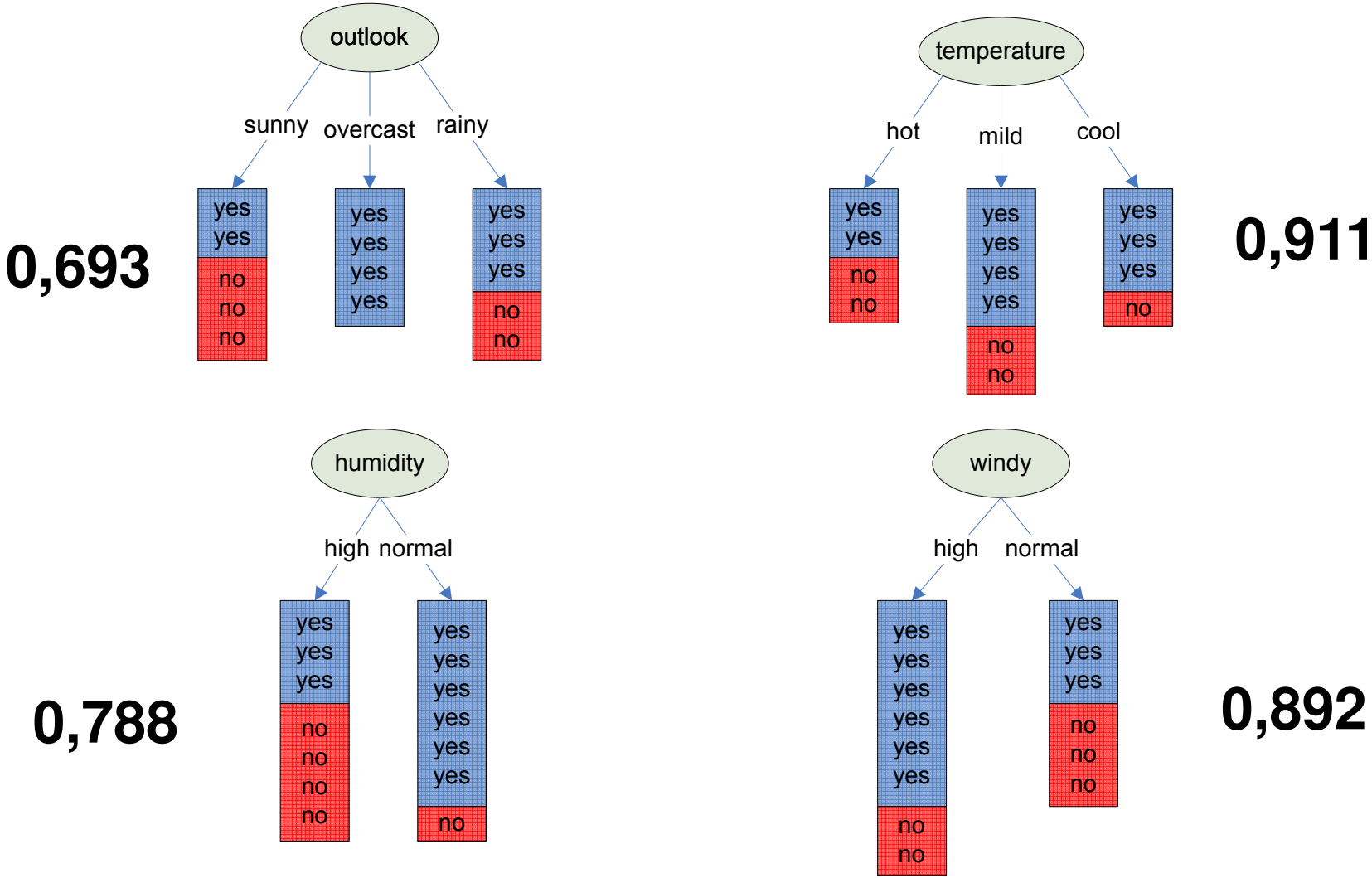
- Welche Frage stellt der Arzt / Kfz-Mechaniker / PC-Experte / Kundenberater als **erste**
  (, um möglichst schnell zum Ziel zu kommen)?

- Erinnerung:
  - Ziel: möglichst kleiner Baum
  - Bei Knoten mit Beispielen einer Klasse (reiner Knoten) ist die Zerlegung beendet

=> **Gierige Heuristik:**

Wähle das Attribut, das die “reinsten” Knoten erzeugt

Unreinheit der Zerlegung E(H)



Das beste Attribut – eine Variante (**ID3**)

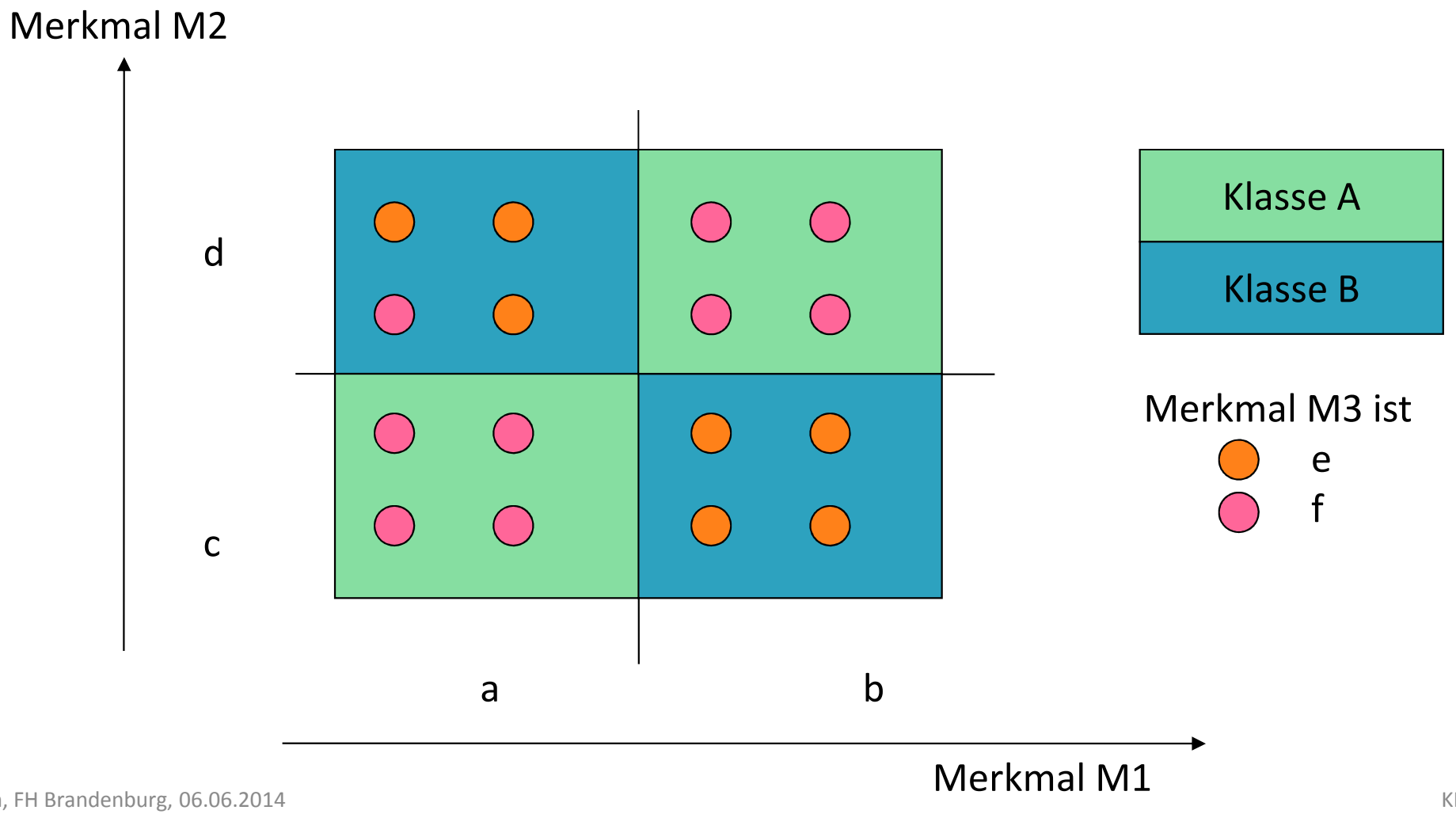
- QUINLAN, 1979 (online in [Qui86])
- nutzt Entropiebegriff von Shannon
- Informationsgewinn (information gain) zur Auswahl des besten Attributs

- Attribut soll möglichst reine Konten erzeugen
- Ein Knoten ist umso reiner, je weniger Fragen zur Klassifikation seiner Beispiele gestellt werden müssen.
  - d. h. je kleiner die noch notwendige Information zur Klassifikation seiner Beispiele ist
  - d. h. je weniger Bits zur Kodierung seiner Beispielmenge mindestens notwendig sind
  - **d. h. je kleiner die Entropie seiner Beispielmenge ist.**

Die Attributauswahl nach Informationsgewinn ist eine Heuristik. Wir hoffen, dass die Wahl des Attributes mit dem höchsten Informationsgewinn dazu führt, dass wir anschließend nur noch wenige Fragen benötigen – das wird in der Regel so sein, aber eben nicht immer.

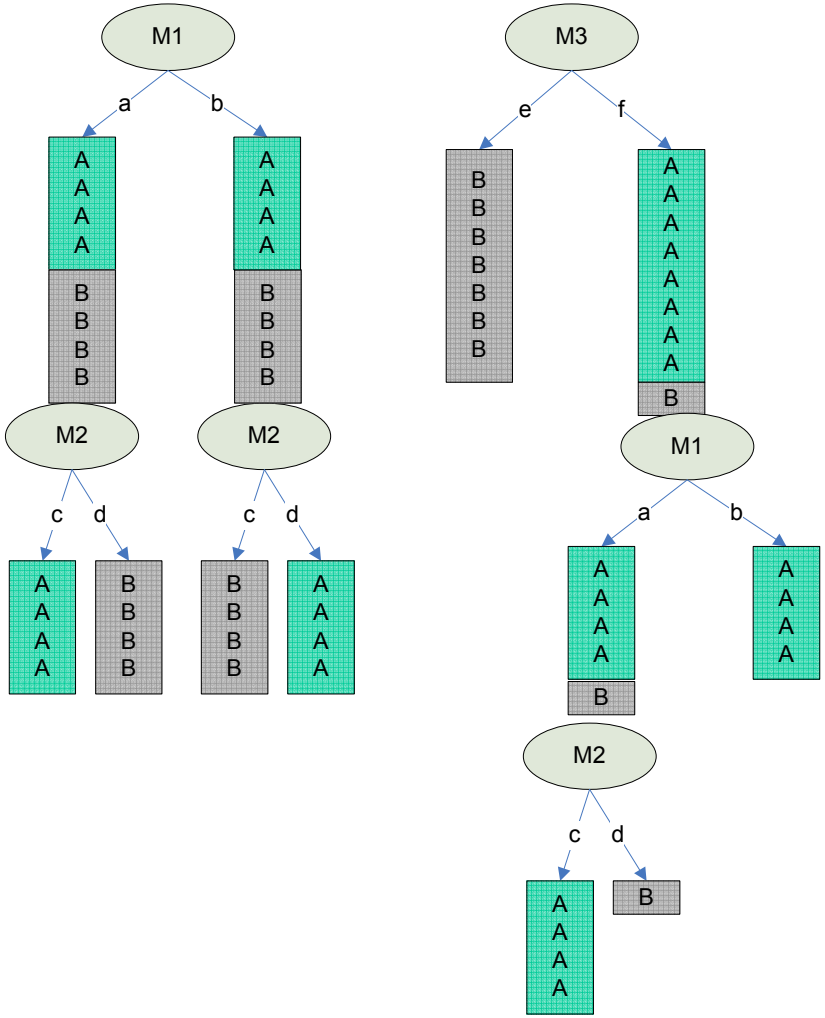
Konstruiertes Gegenbeispiel: Wenn der Informationsgewinn in die Irre lockt

Die Aufteilung nach dem scheinbar günstigem Merkmal M3 (dann M1, M2) erzeugt einen größeren Baum als nach M1, M2.



Gegenbeispiel: Die Bäume

Günstiger Baum nach M1

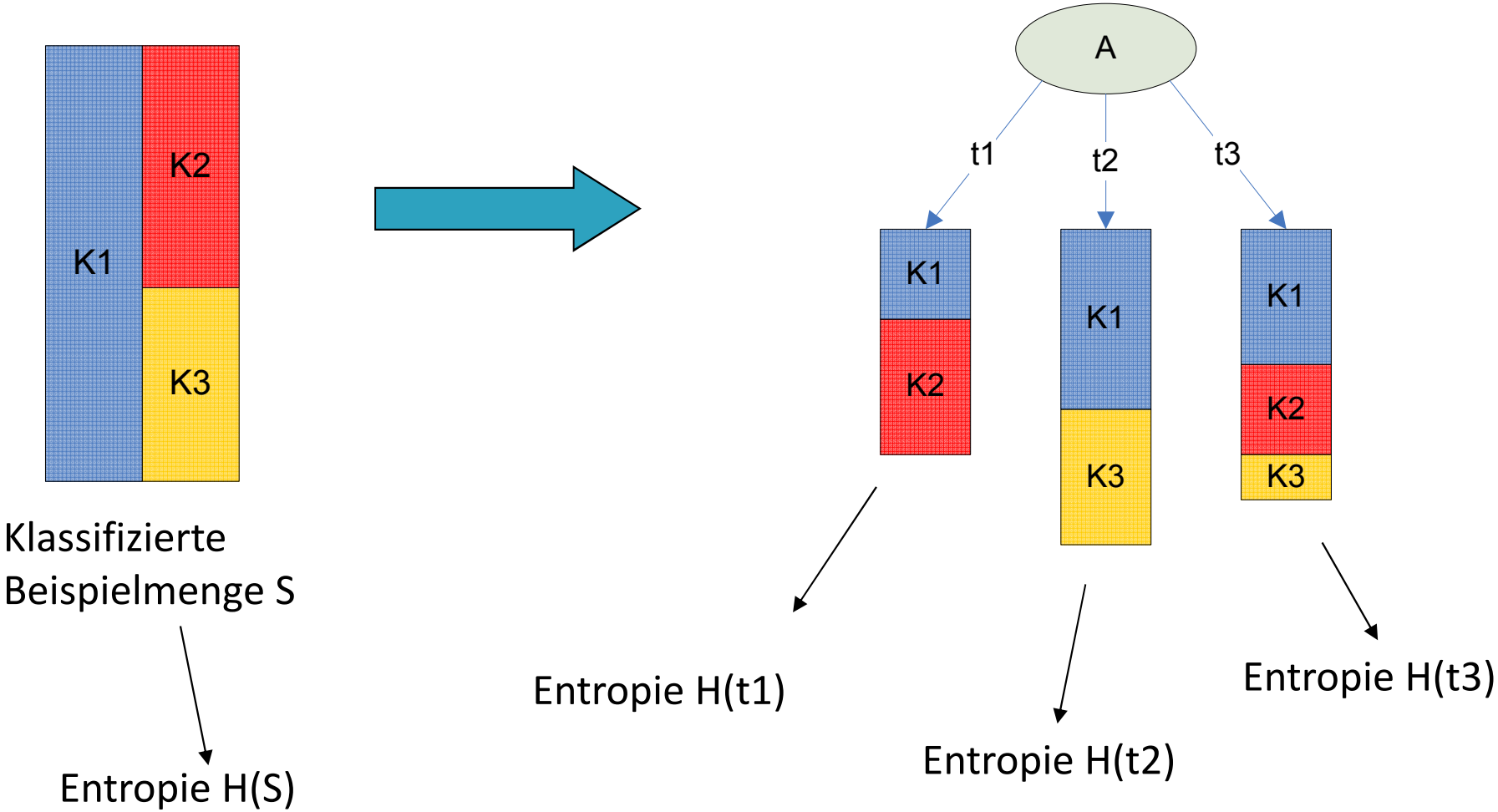


Ungünstiger Baum nach M3

Lassen wir jedoch kleine Fehler in den Knoten zu, so ist die erste Stufe des M3-Baumes doch eine hervorragende Faustregel

ID3 ist ein Bergsteiger-Algorithmus, der gierig und ohne Rückkehr dem Informationsgehalt (geschätzte Kosten zum Ziel) folgt.

Entropie von Symbolmengen



K1 .. K3 sind Klassen, A ist ein Attribut, t1 .. t3 sind die Attributwerte von A

Entropie einer Verteilung (Unreinheit)

Bei zwei Klassen:

S: Menge von klassifizierten Beispielen,
z. B. mit den Klassen K = {yes, no}

|A| = Anzahl der Elemente der Menge A

Relative Häufigkeit von yes in S: p(yes) = |yes| / |S|

Relative Häufigkeit von no in S: p(no) = |no| / |S|

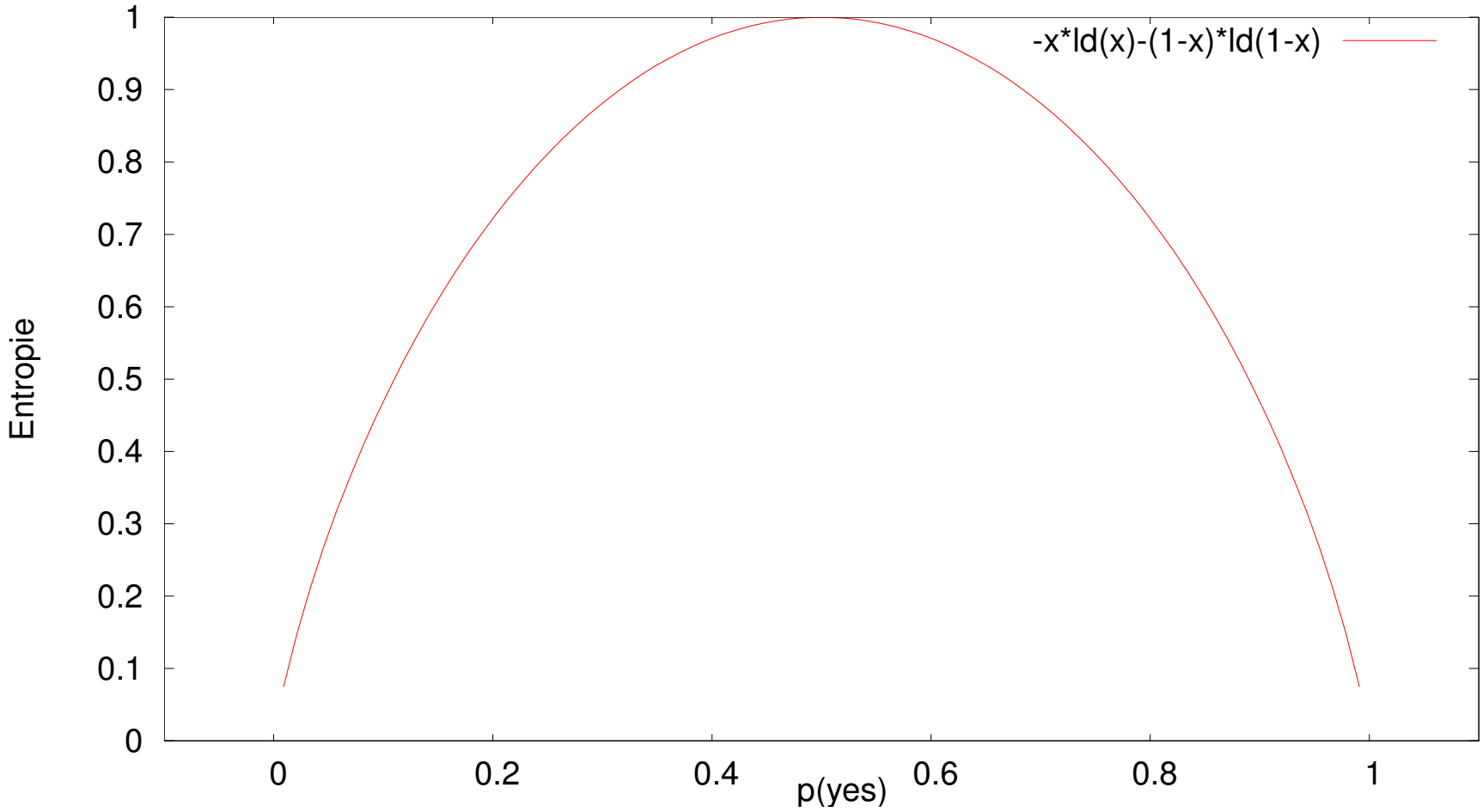
p(yes) + p(no) = 1

Entropie von S?

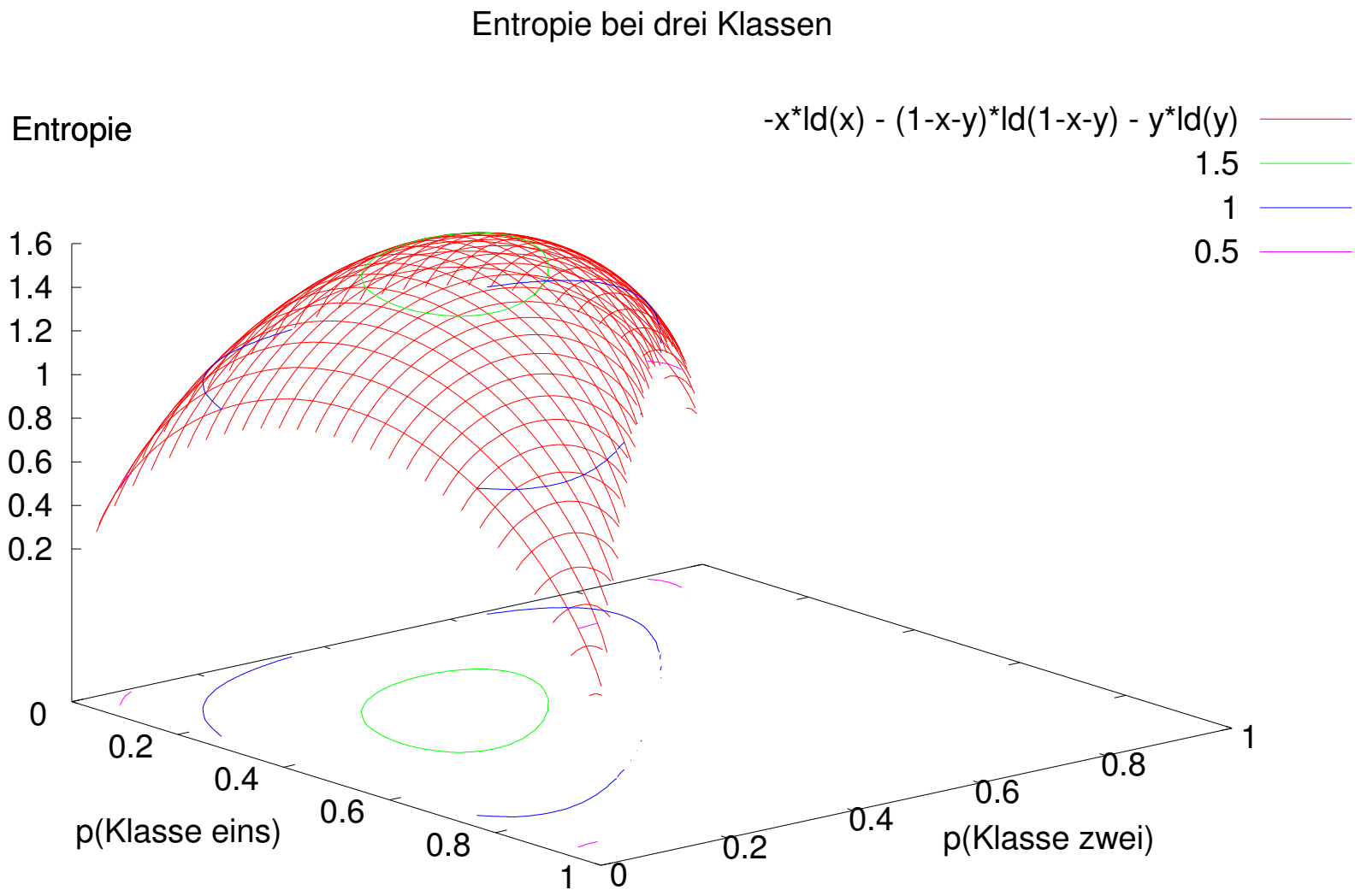
Entropie(S) = H(S) = – p(yes) * ld p(yes) – p(no) * ld p(no)

- H(S) ist am größten, wenn …
- H(S) ist am kleinsten, wenn … oder ....

Entropie bei zwei Klassen



Entropie bei drei Klassen



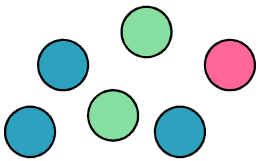
Entropie einer Verteilung (Unreinheit)

Entropie einer Verteilung bei mehreren Klassen:

$$H(S) = -\sum_{j \in K} p(j) \cdot ld\, p(j)$$

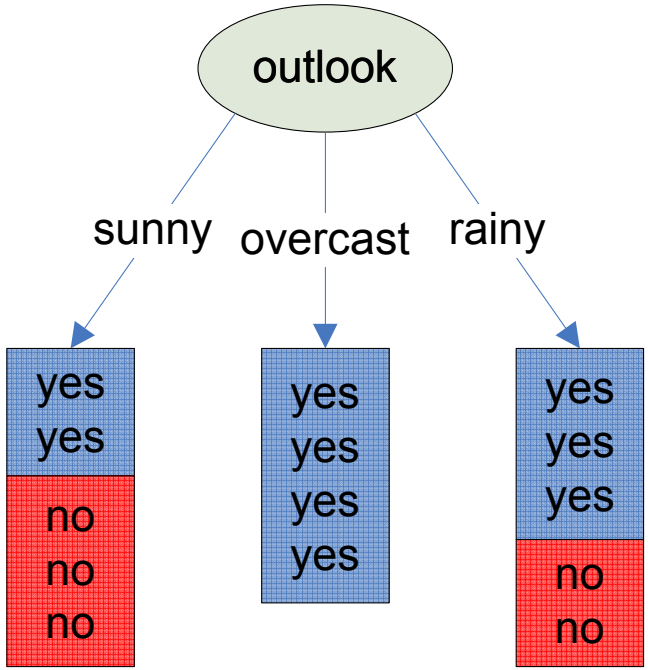
Zwei Schreibweisen:

- H(S) = H(p1, p2, p3 ...) pi sind die **relativen** Häufigkeiten
- H(S) = info(N1, N2, N3 ...) Ni sind die **absoluten** Häufigkeiten

Beispiel: S sei 

H(S) = H(1/6,2/6,3/6) = info(1,2,3)=-1/6*ld1/6-2/6*ld2/6-3/6*ld3/6 = 1.46

Beispiel: Entropie der Aufteilung anhand des Attributes ‚outlook'



Entropie der Ausgangsbeispielmenge S (=vorher)

H(S) =

Entropie der Beispielmenge im Knoten sunny u.a.

H(sunny) =

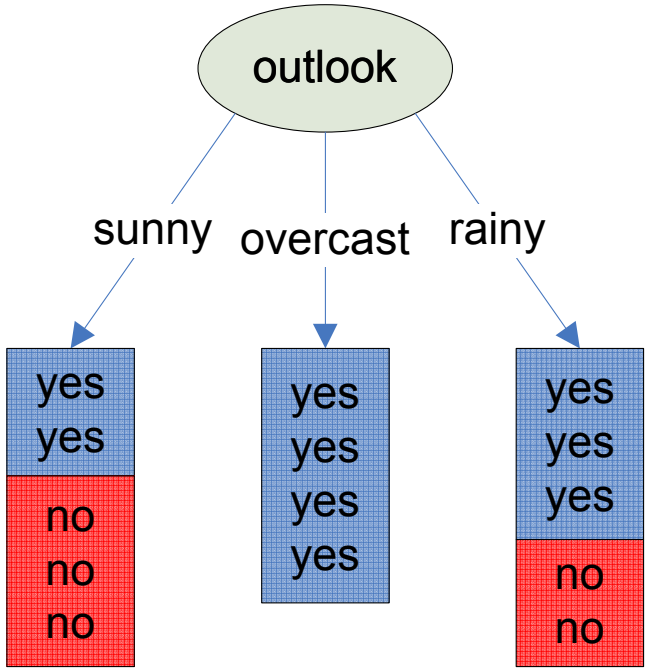
H(overcast) =

H(rainy) =

Entropie von S, sunny, overcast, rainy berechnet – aber wie groß ist nun **Informationsgewinn** von outlook (Entropie Vorher vs. Nachher)?

=> **Erwartete Entropie** (Erwartungswert der Entropie, gewichtete Entropie, mittlere Entropie)

Beispiel: Entropie der Aufteilung anhand des Attributes ‚outlook'



Erinnerung: Erwartungswert einer Zufallsgröße X

$$E(X) = \sum_i p(x_i) \cdot x_i$$

Relative Häufigkeit des Zweiges sunny u.a.

p(sunny) =

p(overcast) =

p(rainy) =

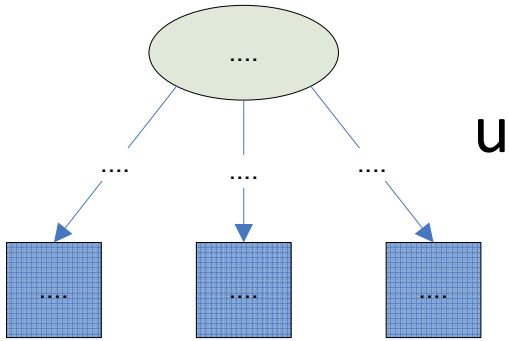
Erwartete Entropie der Zerlegung von S anhand von ‚outlook':

E(outlook) =

p(sunny) * H(sunny) +

p(overcast) * H(overcast) +

p(rainy) * H(rainy) =



Informationsgewinn - Gain

E(outlook)	= 0.693 (noch notwendige Bits)
E(temperature)	= …..
E(humidity)	= …..
E(windy)	= …..
H(S)	= 0.940

vorher *nachher*

Das Kriterium für das beste Attribut:

Höchster Informationsgewinn (Gain) = H(S) – E(Attribut)

Gain(outlook)	= 0.247	=> outlook ist für diese Menge das ‚beste' Attribut
Gain(temperature)	= 0.029	
Gain(humidity)	= 0.152	
Gain(windy)	= 0.048	

Die Bildung der Differenz zu H(S) ist eigentlich nicht mehr nötig, kleinstes E(A) genügt

Zusammenfassung Gain(S,A)

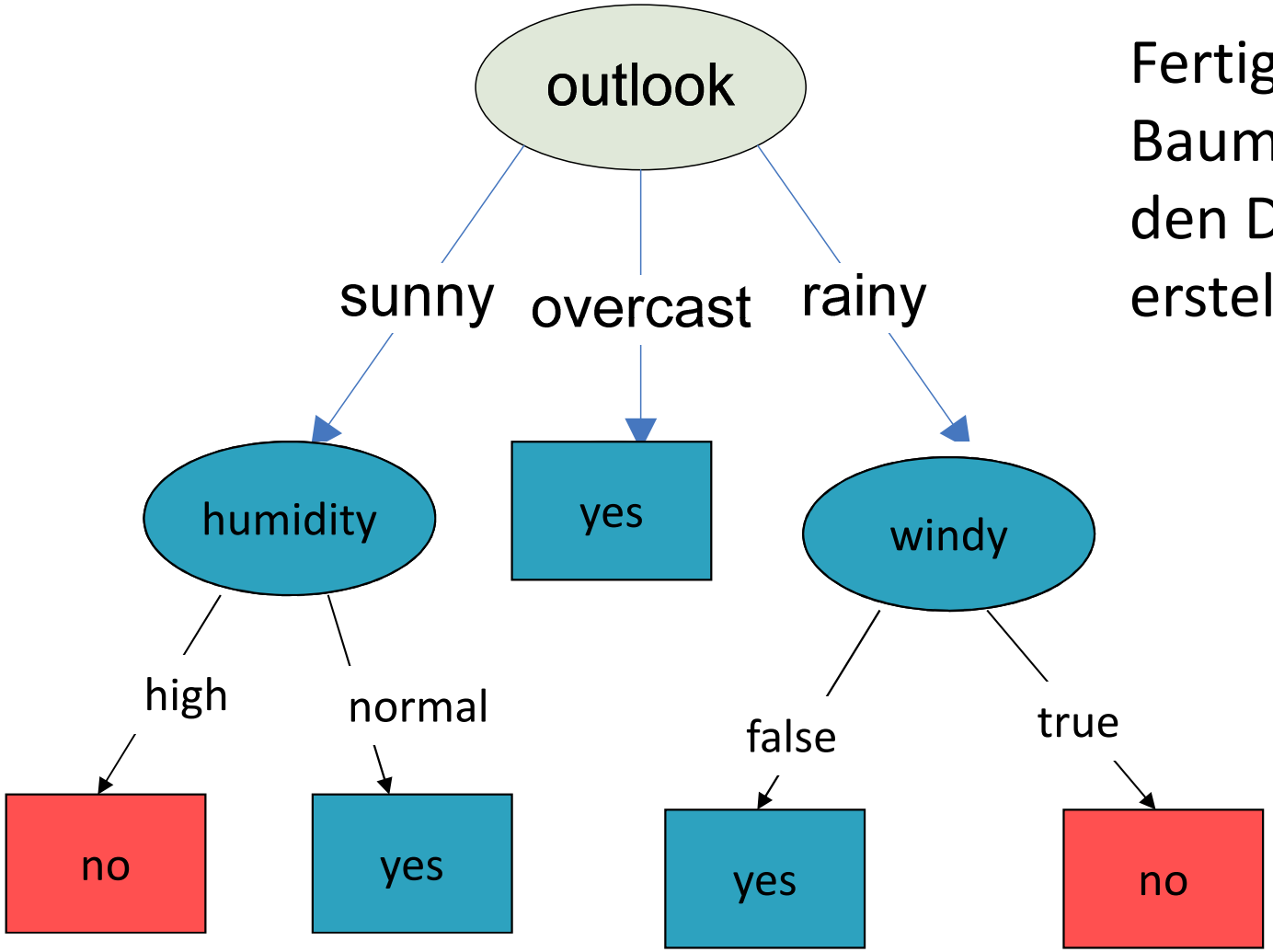
Gain(S,A) = erwartete Verringerung der Entropie nach Zerlegung der Menge S mit dem Attribut A

$$Gain(S,A) = H(S) - \underbrace{\sum_{t \in T(A)} \frac{|S_t|}{|S|} H(S_t)}_{E(A)}$$

- T(A): die möglichen Testergebnisse (Werte) von Attribut A
- S_t: Teilmenge von S, für die das Attribut A den Wert t hat
- H(X): Entropie von X
- E(A): Erwartete Entropie bei Zerlegung von S mit Attribut A

Fragen?

Aufbau des Baumes



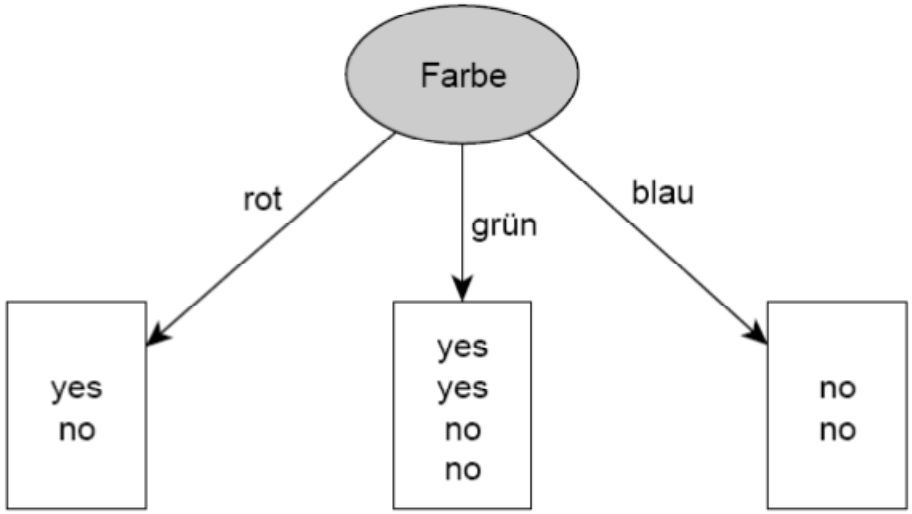
Fertig:
Baum automatisch aus den Datensätzen erstellt

Eine alte Aufgabe

Aufgabe 6: [3+2+1 = 6 Punkte]

a) Mit welchen Aufgabenstellungen befasst sich Data Mining?

b) Ein Attribut *Farbe* teilt die mit *yes* oder *no* klassifizierten Datensätze wie folgt auf:



Bestimmen Sie die Entropien der drei Attributwerte.

c) Wie groß ist die erwartete Entropie des Attributes *Farbe*?

Next

- DM und KDD, Phasen
- Aufgabenstellungen des DM
- Wissensrepräsentation
- Entscheidungsbäume I – Repräsentation
- Entscheidungsbäume II – Lernen

- **Entscheidungsbäume III – Praktisch**
  - Ausblick ID3 zu C4.5
  - Ein Problem
  - Tools
  - RapidMiner
- Performance von Klassifikatoren
- Ethik